UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

TRABAJO FIN DE MÁSTER

Creación de un LIMS para NextSeq

Máster Universitario en BioInformática y Biología Computacional

Autor: Chapado Chorro, Luis Antonio

Tutor: Cuesta de la Plaza, Isabel

Departamento de Ing Informática

FECHA: Febrero, 2018

Contenido

1 Resumen

La tecnología de NGS (Next Generation Sequencing) ha cambiado enormemente nuestro conocimiento de los procesos moleculares proporcionando una ingente cantidad de datos a precios asequibles. Sin embargo, manejar y analizar una cantidad de datos tan grande precisa de unas nuevas herramientas informáticas que ayuden a los profesionales con este nuevo desafío.

iSkyLIMS nace con el objetivo de ayudar en las tareas de laboratorio, implementando un proceso de registro y seguimiento de carreras, muestras y proyectos, reemplazando gran parte de actividades que se estaban realizando de forma manual por procesos automáticos, consiguiendo con ello reducir el número de errores.

iSkyLIMS está orientado a trabajar con el secuenciador de Illumina NextSeq y la plataforma de BaseSpace para optimizar el transvase de información, mostrando no sólo la información de la carrera, sino siendo un entorno donde se pueden hacer consultas de las carreras realizadas, relacionando entre sí las carreras, proyectos, muestras e investigadores. La plataforma implementa una serie de estadísticas e informes que van a permitir detectar fallos de calidad en los distintos pasos del proceso, e introducir acciones correctivas.

iSkyLIMS se ha diseñado para que pueda funcionar en un entorno virtual, siendo de esta forma escalable y facilitando su uso al investigador permitiendo una optimización de las tareas de laboratorio.

2 Objetivos

Los objetivos que se plantea iSkyLIMS son:

- Desarrollar una plataforma Web para la incorporación de la gestión de datos y muestras en la Unidad de Genómica.
- La identificación y seguimiento de las muestras.
- Almacenamiento en una base de datos, de la información de calidad de los proyectos y carreras y su asociación con los usuarios.
- Implementación y visualización de estadísticas por carrera, proyectos, muestras, usuarios y kits de secuenciación.

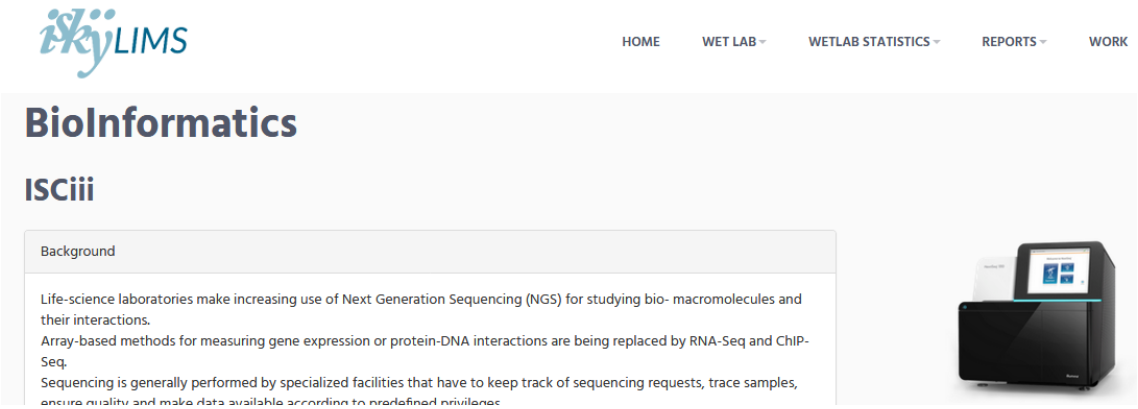


Ilustración 1 Página principal de iSkyLIMS

3 Introducción

En la actualidad muchos laboratorios que trabajan con muestras genómicas se enfrentan al problema del seguimiento de muestras, y donde la incorporación de un software LIMS (Laboratory Information Management System) suele ser la solución para asegurar el correcto manejo de las muestras.

3.1 ¿Qué es un LIMS?

El término LIMS proviene de las siglas en ingles de Laboratory Information Management System. Se basa en un Software que maneja la información que se produce en el laboratorio, que contiene la funcionalidad necesaria para realizar las operaciones diarias de un laboratorio en el que se gestionan una gran cantidad de datos y que puede estar sometido a un proceso de control de calidad.

Hoy en día la funcionalidad de un LIMS no se limita a tener un flujo de trabajo y un control de seguimiento de las muestras, sino que las funcionalidades están evolucionando para adaptarse a las nuevas necesidades como son; el intercambio de datos entre distintos interfaces y ser implementadas sobre arquitecturas más flexibles y eficientes.

3.2 LIMS existentes

En la actualidad hay una gran variedad de LIMS, que pueden obtenerse a través de:

- Marcas comerciales como Illumina, ofreciendo productos como:
  - NGS LIMS, Illumina Array LIMS, Clarity LIMS
- Empresas de servicios de Software :
  - Nevis LIMS, LabWare LIMS, STARTLIMS
- LIMS open source.
  - Bika LIMS, GNomEX, FreeLIMS, Open-LIMS

De todos ellos hemos elegido aquellos que son open source y que se mencionan en la siguiente tabla.

Nombre	Referencia	Descripción
MendeLIMS	MendeLIMS: a web-based laboratory information management system for clinical genome sequencing	LIMS para el manejo de muestras de NGS para la identificación de patologías en pacientes
SLIMS	SLIMS—a user-friendly sample operations and inventory management system for genotyping labs	LIMS que gestiona la información de los pacientes, las muestras biológicas y los contenedores utilizado para almacenar y enviar.
SMITH	SMITH: a LIMS for handling next-generation sequencing workflows	LIMS para el analysis de los experimentos basados en Chip-Seq, RNA-seq, miRNA-Seq y Exome-Seq
Galaxy LIMS	Galaxy LIMS for next-generation sequencing	LIMS optimizado para trabajar con illumina Hi Seq 2000
MetaLIMS	MetaLIMS, a simple open-source laboratory information management system for small metagenomic labs	LIMS para laboratorios pequeños de metagenómica que desean almacenar colecciones de muestra y el procesamiento de la información, pero no necesitan el volumen adicional de grabando NGS o datos de análisis
adLIMS	adLIMS: a customized open source software that allows bridging clinical and basic molecular research studies	LIMS para evitar el uso de hojas de cálculo o archivos locales utilizando para ello procedimientos estándar para el seguimiento de muestras

Tabla 1 Referencia de LIMS open source

3.3 ¿Por qué iSkyLIMS?

Después de ver la cantidad de LIMS que están accesibles para un laboratorio, nos cabría preguntar ¿Por qué hacer uno más? ¿No hay ninguno que satisfaga nuestras necesidades?

Desgraciadamente, la respuesta a esta pregunta es que ninguno de ellos está manejando el equipo de NextSeq de illumina y que el coste de adaptar alguno de los LIMS existentes es similar o mayor al de crear un LIMS nuevo adhoc orientado y optimizado a funcionar en el entorno y requerimientos del ISCIII.

Por esta razón se ha creado iSkyLIMS para poder realizar los siguientes objetivos:

- Manejo de la demandante explosión de Información que supone una carrera de secuenciación masiva.
- Asegurar la calidad de las muestras.
- Reducción de la tasa de errores de entrada de datos
- Necesidad de una respuesta más rápida en la relación muestra/ tiempo de los resultados
- Reducir la dependencia con la plataforma de BaseSpace de illumina.

El uso de la plataforma de BaseSpace está basada en créditos. Cuando un usuario abre una cuenta en BaseSpace se le otorga una serie de créditos que va gastando al utilizar los servicios que ofrece la plataforma. Una vez gastados es necesario comprar nuevos créditos. Adicionalmente a este coste, los datos del investigador, son automáticamente propiedad de illumina, tal y como se refleja en las condiciones de aceptación del servicio a BaseSpace.

> 7. The information, data, documents, tools including design tools, software, and other materials accessible through MyIllumina ("Materials") are proprietary to Illumina, its affiliates, licensors, or other third parties.

Aunque cualquiera de estas razones, son suficientes para hacernos pensar en la necesidad de tener un sistema de datos centralizados, va a ser de obligado cumplimiento a la hora de obtener cualquier certificación (por ejemplo, la ISO 9001).

Por ello se necesita estandarizar los sistemas de administración y seguimiento de datos, construyendo una estructura escalable y flexible con interfaces basadas en web, que generalmente se llaman Sistema de gestión de información de laboratorio (LIMS).

La gestión del laboratorio y de sus herramientas han de estar totalmente unidas para ofrecer la eficiencia y los óptimos resultados, por ello no se puede utilizar cualquier LIMS existentes sin evaluar si va a cumplir con los requerimientos definidos en el laboratorio. Aun así, muy posiblemente haya que adaptar y/o crear numerosos scripts para que realmente pueda ser usado por el laboratorio.

iSkyLIMS va a llevar el control de la gestión de las muestras de NextSeq creadas en el laboratorio del iSCIII, sacar datos y gráficas de las carreras, sacar estadísticas relacionadas con la gestión del laboratorio, y la obtención de informes mensuales y anuales.

Estos informes y estadísticas van a ser de gran ayuda para poder determinar la calidad con la que se están generando las muestras en el iSCIII, permitiendo de esta forma crear acciones para mejorar la calidad de las muestras y tener un medio para poder cuantificarlas.

4 Materiales y Métodos

4.1 Infraestructura Hardware

El ISCIII cuenta con una plataforma de computación de alto rendimiento (High Performance Computing, HPC) con 320 cores, 4096 GB memoria RAM, donde se realiza la conversión de los ficheros bcl generados en la secuenciación a ficheros fastq que contienen las lecturas con la calidad asociada a cada base. Además, cuenta con una cabina de almacenamiento NetApp de 70TB netos donde se almacenan los datos provenientes del secuenciador.

iSkyLIMS ha sido desarrollado en una máquina virtual desplegada sobre VMCenterv6.0 en un servidor ubicado en el Centro de Procesamiento de Datos del ISCIII (CPD) (HP Proliant D385 G7, AMD Opteron de 16 núcleos (x2), 128GB RAM, 8TB tipo SAS). La

máquina corre el sistema operativo Centos 6.9. donde está configurado el entorno de desarrollo y un servidor web Apache con módulo WGSI que sirve la aplicación, y un servidor Nginx para servir los ficheros web estáticos.

4.2 Recogida de requerimientos

El método que se ha utilizado es el recoger requerimientos de los investigadores, en términos de la funcionalidad, accesibilidad y entorno gráfico que debería tener la herramienta. Así mismo se recogieron los requerimientos de la unidad de seguridad estableciendo los criterios que debía de cumplir. Por otro lado, la herramienta debía formar parte del entorno del departamento de Bioinformática, integrándose con el entorno de producción existente y encajar dentro de la estrategia de evolución del departamento.

Los principales requerimientos fueron:

- Eliminación de la duplicidad a la hora de crear el Sample Sheet
- Creación de un nuevo flujo de usuario que evitase el envío de información fuera del ámbito del ISCII
- Conexión con el servidor de almacenamiento para coger los datos de las carreras mediante SAMBA
- Entorno gráfico basado en WEB, usando formularios para la creación de las carreras
- Obtener y mostrar estadísticas con la calidad de las carreras, de los proyectos y de las librerías usadas
- Estar integrado con el HPC de la unidad de Bioinformática
- Restricción de acceso a la página Web en función del tipo de usuario

4.3 Entorno de desarrollo

ISkyLIMS es un software open source (https://github.com/BU-ISCIII/iSkyLIMS-wetlab)está implementado en un entorno web utilizando el framework de desarrollo Django, versión 2.0.1, (un framework desarrollado en Python) para procesar las peticiones del usuario y mostrar los resultados.

Django está corriendo sobre Python versión 3.6.1, utilizando paquetes de Python para facilitar el procesamiento del código, como son; BioSeq, interop, ElementTree, entre otros.

La parte frontend está desarrollada en Java Script, HTML5 y css. En particular usamos el framework jquery de javascript y la hoja de estilos bootstrap para la presentación de la información de la página Web.

La información que procesa iSkyLIMS se almacena en una base de datos, usando MySQL versión 15.1.

4.4 Datos de carreras NextSeq

Los datos utilizados para la representación de las gráficas se han basados en los ficheros generados por illumina NextSeq, donde 10 investigadores han utilizado 2 conjuntos de librerías (Nextera XT y TrueSeq Rapid Exome) generando un total de 22 carreras que van desde junio del 2.016 hasta noviembre del 2.017.

4.5 Programas de terceros utilizados

Illumina Experience Manager (IEM), versión 1.14.0.162, se va a utilizar para la colocación de las muestras en el kit de librería y con ello poder crear el Sample Sheet.

Illumina Sequecing Analysis Viewer, versión 1.8.37, se ha utilizado para verificar los valores obtenidos por el código de iSkyLIMS son correctos.

Para la conversión de ficheros BCL, generados por el secuenciador de illumnina, al formato fastq, vamos a utilizar el programa bcl2fastq versión 1.8.4.

4.6 Ficheros de Entrada

Los ficheros que va a procesar iSkyLIMS los podemos diferenciar dependiendo del uso que vamos a hacer de ellos que son:

- Definición de la carrera en iSkyLIMS.
- Configuración de la carrera de NextSeq en plataforma BaseSpace.
- Datos de calidad para la realización de estadísticas.

4.6.1 Ficheros para la definición de la carrera en iSkyLIMS

La estructura de los directorios donde se van a localizar los ficheros de entrada van a estar definidos por illumina según se muestra en la figura de abajo.

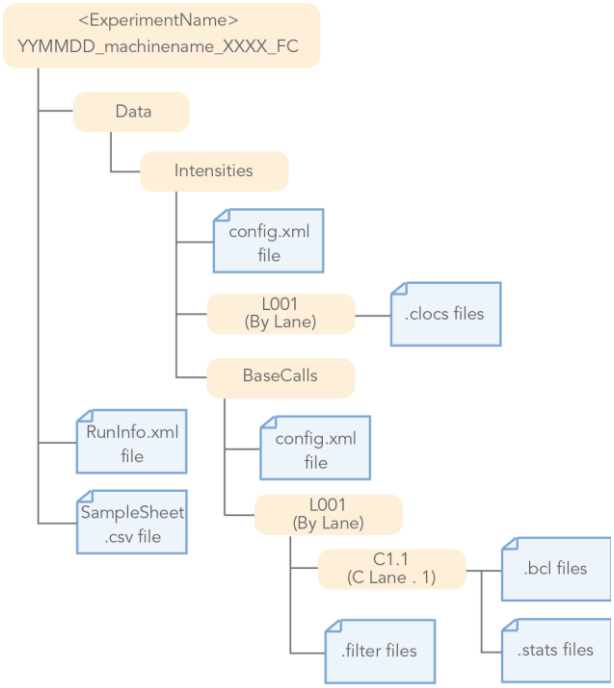


Ilustración 2 Directorios creados por illumina NextSeq

Para empezar a definir la carrera en iSkyLIMS, se necesita como entrada del formulario el fichero **SampleSheet.csv**.

Este fichero es generado por el programa illumina Experience Manager (IEM) se compone de varias secciones; Header, Reads, Settings y Data con el siguiente formato.

[Header]

Compuesto por los siguientes campos:

IEMFileVersion	**5**
Date	**05/11/2017**
Workflow	**GenerateFASTQ**
Application	**NextSeq FASTQ Only**
Instrument Type	**NextSeq/MiniSeq**
Assay	**Nextera XT**
Index Adapters	**Nextera XT Index Kit (24 Indexes, 96 Samples)**
Description	**descripción del proyecto**
Chemistry	**Amplicon**

Tabla 2 Campos incluidos en la sección de Header en SampleSheet.csv

[Reads]

- Va a contener el número de las lecturas utilizadas

[Settings]

Compuesto por el adaptador

Adapter	CTGTCTCTTATACACATCT

Tabla 3 Campo definido en la sección de Settings de SampleSheet.csv

[Data]

Esta sección está compuesta por una cabecera, según se muestra en figura siguiente.

Sample_ID	Sample_Name	Sample_Plate	Sample_Well	I7_Index_ID	index	I5_Index_ID	index2	Sample_Project	Description

Ilustración 3 Campos incluidos en la sección de Data en ShampleSheet.csv

Donde las columnas de cada fila van a contener el identificador de la muestra y su información descrita en la cabecera.

4.6.2 Ficheros de configuración de la carrera

Dos ficheros son creados por el secuenciador para indicar los datos con los que se ha realizado la carrera.

- RunInfo.xml
- RunParameters.xml

De ellos obtendremos la información de; el identificador de la carrera que ha generado el secuenciador, los canales de imagen utilizados, el Flowcel, las versiones utilizadas en la “suite” y en la “RTA”

Otro fichero que vamos a utilizar es:

- RunCompletionStatus.xml

De este fichero solo vamos a comprobar la línea “CompletionStatus” para identificar si la secuenciación se ha completado correctamente o si por el contrario se ha abortado.

En el caso de que se haya parado la secuenciación por la petición del “Wetlab manager” o por problemas técnicos, la carrera aparecerá en estado de “error”.

4.6.3 Ficheros de calidad para la realización de estadísticas

La mayoría de los ficheros que iSkyLIMS que necesita procesar van a provenir del proceso de conversión de BCL a Fastq, donde el programa creará la siguiente estructura de directorios.

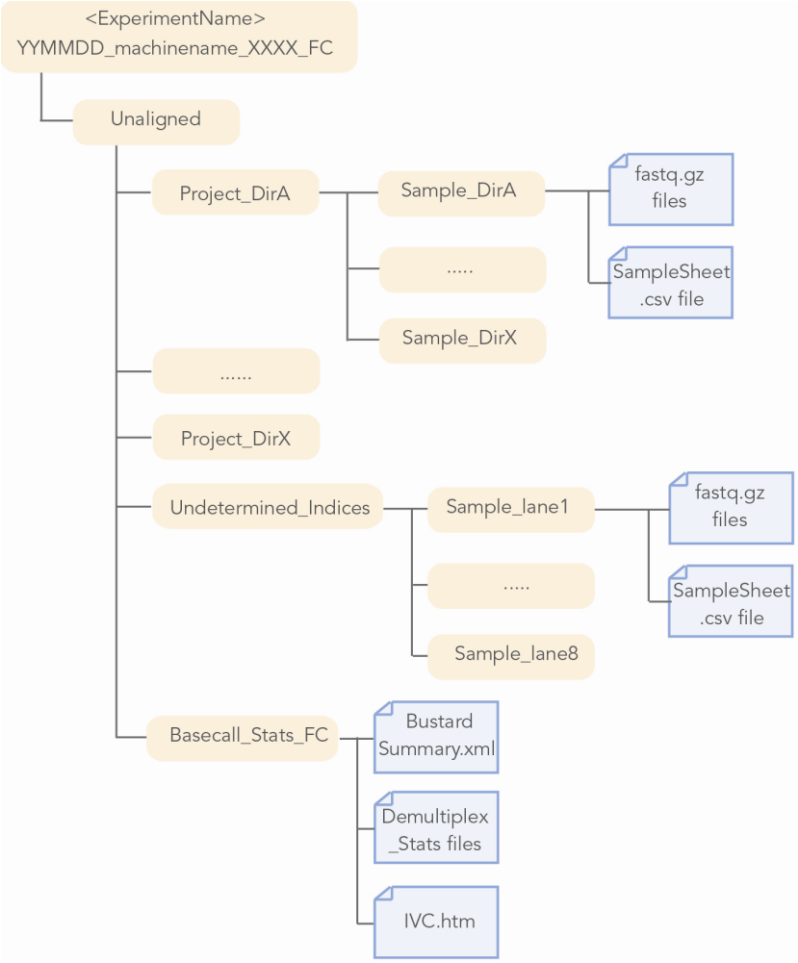


Ilustración 4 Directorios creados después de la ejecución de la conversión bcl2fastq

Dentro del conjunto de ficheros que vamos a utilizar para este propósito, se pueden clasificar en 2 grupos:

- Información de la calidad de las muestras
- Información de la calidad de la carrera.

Para obtener información de la calidad de las muestras usaremos los ficheros:

- ConversionStats.xml
- DemultiplexingStats.xml

Donde obtendremos el número de bases de cada muestra, el porcentaje de Barcode, número de clústeres, número de muestras con una calidad mayor o igual a 30, y la media de la calidad de las muestras.

El otro grupo de ficheros que obtendremos de la conversión de BCL a fastq son los que se crean en la carpeta de “interop” y que van a darnos información de la calidad de la carrera

- CorrectedIntMetricsOut.bin
- ExtractionMetricsOut.bin
- RegistrationMetricsOut.bin
- EmpiricalPhasingMetricsOut.bin
- IndexMetricsOut.bin

- TileMetricsOut.bin
- ErrorMetricsOut.bin
- PFGridMetricsOut.bin
- EventMetricsOut.bin
- QMetricsOut.bin

Obteniendo valores de alineamiento, tasa de errores, número de muestras con una calidad mayor o igual a 30, número de bases secuenciadas.

4.7 Conversión bcl2fastq y compartición de datos con usuario

En el flujo de usuario comentamos que se iniciaba con la preparación de las muestras y terminaba una vez que el wetlab manager ordenaba.

A partir de este punto se realizará la secuenciación dando como resultado unos ficheros de imágenes que serán enviados tanto a la plataforma de BaseSpace como a la cabina de almacenamiento asociada al HPCdel ISCIII. Un proceso automatizado implementado en el entorno HPCse encargará de ir comprobando periódicamente si se ha finalizado la transferencia de los ficheros “bcl”. Cuando esté finalizada este proceso verificará si está el fichero Shample Sheet en el directorio. Cuando estas 2 condiciones se cumplen se ejecutará el comando bcl2fastq para la conversión al formato fastq y los ficheros crudos se compartirán vía samba (dado que su entorno es Windows) a los usuarios, que recibirán una notificación por mail para su descarga

Durante esta conversión se van a generan además ficheros que calidad y medidas de las muestras que serán utilizadas posteriormente por iSkyLIMS.

5 Resultados

5.1 1. Visión general del entorno iSkyLIMS

Para poder interactuar con los datos generados por el secuenciador de illumina NextSeq, la aplicación, integrada dentro de iSkyLIMS, ha de poder conectarse al servidor donde se van a almacenar todos los ficheros que el secuenciador genera. Es decir que el servidor donde está corriendo iSkyLIMS ha de tener comunicación con la cabina de almacenamiento para poder acceder a los ficheros generados en la carrera.

La figura de abajo muestra la conexión lógica entre los diferentes nodos involucrados en la solución de iSkyLIMS.

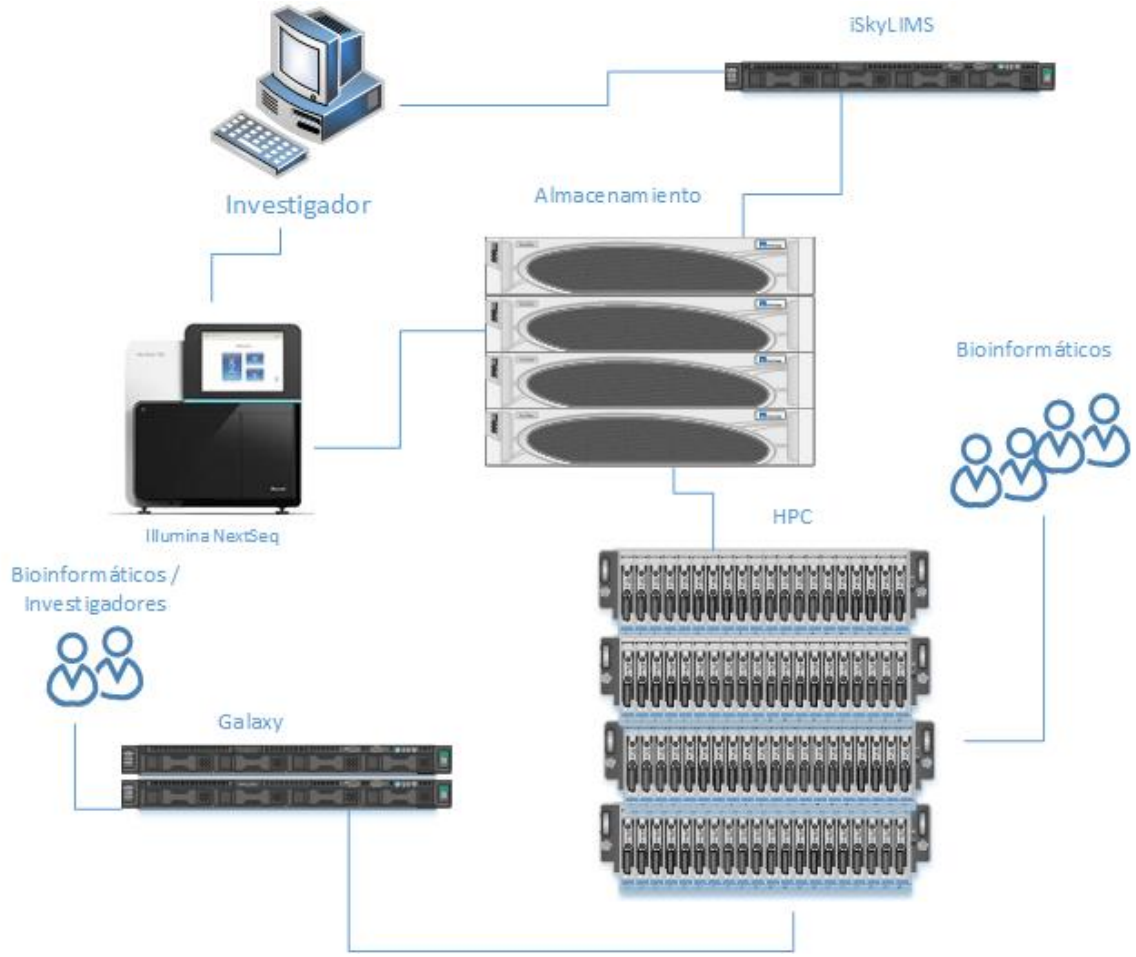


Ilustración 5 Estructura lógica del entorno de iSkyLIMS

En la figura vemos que el Wetlab manager se conectará a iSkyLIMS para poder realizar las tareas de creación de carrera o bien para obtener información estadísticas o informes.

Para la realización de estas estadísticas iSkyLIMS se ha tenido que conectarte anteriormente al servidor de almacenamiento donde están almacenados los datos que suministra la plataforma de illumina y donde se guardan también los ficheros generados por la ejecución, en el Cluster de Alta Disponibilidad (HPC), de la conversión de bcl2fastq.

Este cluster está siendo compartido para realizar los trabajos de los bioinfomáticos y de las herramientas que son lanzadas para su ejecución desde los servidores de Galaxy.

5.2 Funcionamiento de iSkyLIMS

Para explicar el funcionamiento de iSkyLIMS, vamos a dividirlos en los siguientes apartados:

- Roles de usuarios
- Flujo de usuario

- Evolución de estados de la carrera
- Estructura de la base de datos
- Logs
- Copias de seguridad

5.3 Roles de usuarios

Se han definido 4 tipos de usuarios dependiendo de sus roles:

- Investigador
- Investigador Leader
- Wetlab Manager
- Administrador

Cada tipo de usuario tiene acceso a un conjunto de información acorde a su rol.Para acceder a iSkyLIMS será necesario acceder con un nombre de usuario y contraseña.Una vez que se accede iSkyLIMS identificará al usuario dentro de uno de estos 3 roles permitiéndole sólo a las páginas autorizadas para su rol.

5.3.1 Rol de Investigador

El investigador va a ser la persona que se va a encargar de preparar las muestras del proyecto. Este roltendrá acceso a la búsqueda de los proyectos en los que ha participado.

5.3.2 Rol de Investigador Leader

El investigador Leader va a ser la persona que va supervisar la calidad del proceso de preparación de las muestras del proyecto. Este rol tendrá acceso a la búsqueda de todos los proyectos.

5.3.3 Rol de Wetlab manager

El Wetlab Manager va a ser el encargado de la creación del Sample Sheet y la de interactuar con iSkyLIMS para la definición de la carrera. Será la persona que ordene la ejecución de la secuenciación en el secuenciador de illumina.Tendrá acceso toda la parte de Wetlab de iSkyLIMS, pudiendo acceder a las estadísticas, informes y datos de las carreras y proyectos de todos los investigadores.

5.3.4 Administrador

El rol de administrador le permitirá el acceso total a iSkyLIMS, permitiéndole incluso a definir una carrera en iSkyLIMS.Además, tendrá acceso a la parte de administración de iSkyLIMS otorgando y cambiando permisos de usuarios.

5.4 Flujo de Usuario

En esta sección se muestra las acciones que debe realizar el Wetlab Manager para realizar la secuenciación de la carrera. Este flujo lo podemos dividir en las siguientes partes:

- Preparación de la muestra
- La que se realiza dentro iSkyLIMS
- La realizada utilizando la plataforma de illumina (Base Space)

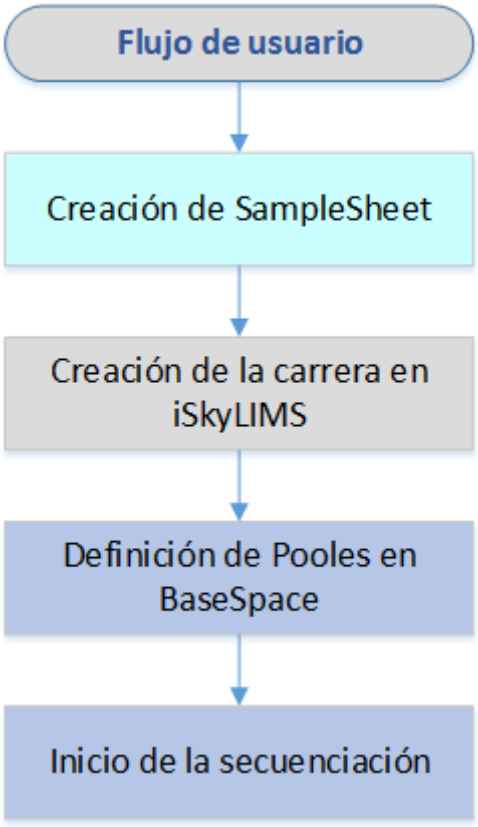


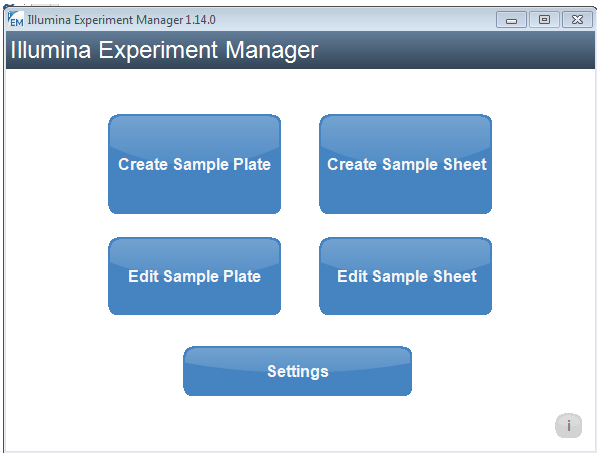
Ilustración 6 Control de flujo del usuario para la creación de una carrera en iSkyLIMS

5.4.1 Creación del fichero “SampleSheet”

Para la creación del fichero “SampleSheet.csv” vamos a utilizar el software Illumina Experiment Manager (IEM). Este software proporcionado por illumina, nos va a servir de ayuda para crear y editar hojas de muestra, “Sample Sheet.

La recomendación que desde Illumina nos aconsejar seguir es la de usar el Administrador de experimentos antes de comenzar la preparación de la muestra o la biblioteca, ya que el Administrador de experimentos puede detectar y advertir combinaciones de índices subóptimos y de combinaciones no posibles.

La creación de un Sample Sheet se compone de 2 pasos:



1. **Crear una placa de muestra:** En este paso, se almacena información sobre las muestras en cada pocillo de una placa. Esta información incluye el tipo de preparación de la biblioteca que se realiza, el nombre de la placa y los índices de muestra.
2. **Crear una hoja de muestra:** Que se basa en la información definida durante la creación de la placa de muestra.

Ilustración 7 Ventana principal de illumina Experiement Manager

Para crear la placa de muestra usaremos el asistente que tiene este software para ayudarnos en la confección de la placa.

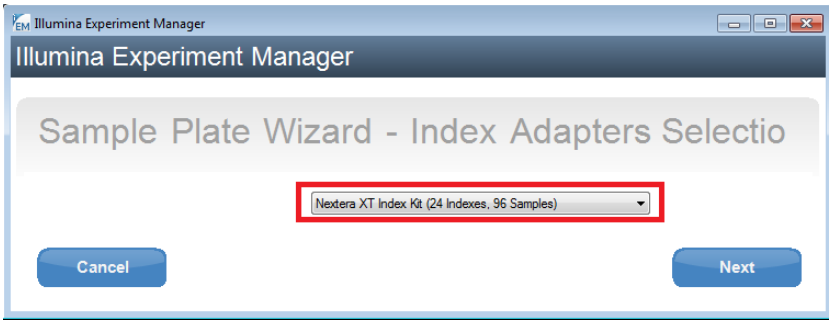
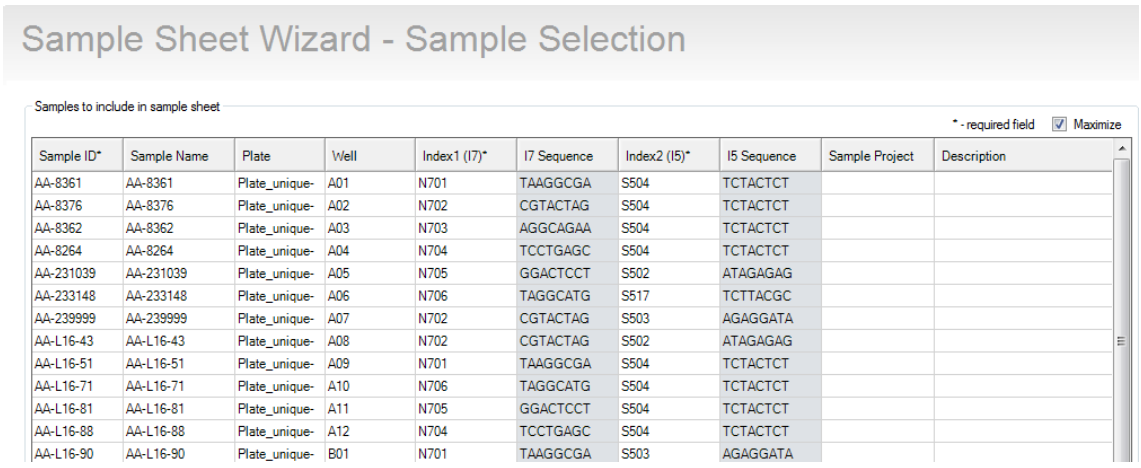


Ilustración 8 Elección del "plate" dentro del illumina Experiement Manager

Una vez que se ha completado utilizaremos el asistente que Sample Sheet que también está incluido en el software.



Sample ID*	Sample Name	Plate	Well	Index1 (I7)*	I7 Sequence	Index2 (I5)*	I5 Sequence	Sample Project	Description
AA-8361	AA-8361	Plate_unique-	A01	N701	TAAGGCGA	S504	TCTACTCT		
AA-8376	AA-8376	Plate_unique-	A02	N702	CGTACTAG	S504	TCTACTCT		
AA-8362	AA-8362	Plate_unique-	A03	N703	AGGCAGAA	S504	TCTACTCT		
AA-8264	AA-8264	Plate_unique-	A04	N704	TCCTGAGC	S504	TCTACTCT		
AA-231039	AA-231039	Plate_unique-	A05	N705	GGACTCCT	S502	ATAGAGAG		
AA-233148	AA-233148	Plate_unique-	A06	N706	TAGGCATG	S517	TCTTACGC		
AA-239999	AA-239999	Plate_unique-	A07	N702	CGTACTAG	S503	AGAGGATA		
AA-L16-43	AA-L16-43	Plate_unique-	A08	N702	CGTACTAG	S502	ATAGAGAG		
AA-L16-51	AA-L16-51	Plate_unique-	A09	N701	TAAGGCGA	S504	TCTACTCT		
AA-L16-71	AA-L16-71	Plate_unique-	A10	N706	TAGGCATG	S504	TCTACTCT		
AA-L16-81	AA-L16-81	Plate_unique-	A11	N705	GGACTCCT	S504	TCTACTCT		
AA-L16-88	AA-L16-88	Plate_unique-	A12	N704	TCCTGAGC	S504	TCTACTCT		
AA-L16-90	AA-L16-90	Plate_unique-	B01	N701	TAAGGCGA	S503	AGAGGATA		

Ilustración 9 Selección de las muestras dentro de illumina Experiement Manager

Tras la verificación de que las muestras están en la posición correcta y que los índices son los apropiados exportaremos esta información a un fichero (SampleSheet) en formato csv, salvándolo en nuestro ordenador. Este fichero será el que se utilizará en el paso siguiente.

5.4.2 Creación de la carrera en iSkyLIMS

La creación de la carrera dentro de iSkyLIMS consta de 2 pasos que son guiados mediante la presentación al usuario del formulario adecuado para que sea rellenado.

El primer formulario se crea para que el usuario introduzca el SampleSheet.csv dentro de la herramienta.

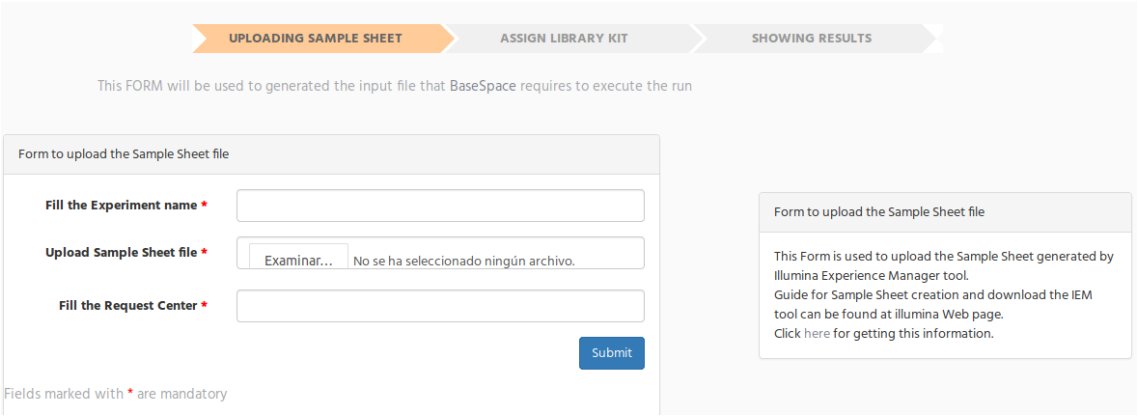


Ilustración 10 Primer paso de la creación de una carrera dentro de iSkyLIMS

Como datos adicionales se debe de introducir el nombre del experimento y el nombre del Centro.

Al dar al botón de “Submit” al usuario le aparecerá un nuevo formulario para poder introducir el nombre del Kit de librería que se ha utilizado en el proyecto.

Al dar de nuevo al botón de “Submit” le aparecerá el fichero que ha de ser importado en el entorno de BaseSpace.

Para descargarlo se hará click en el icono de descarga y se almacenará en el ordenador del usuario para su posterior importación a la web de BaseSpace.

5.4.3 Definición de Pooles en BaseSpace

Una vez que se han obtenido el fichero o los ficheros de las librerías, que se van a utilizar en la carrera, el wetlab Manager los utilizará para la preparación de la librería en Base Space.

El wetlab Manager, usando sus claves de acceso para la página de Base Space (https://basespace.illumina.com/) , se elegirá del menú la opción de preparativos de la librería (PREP) tal y como se muestra en la figura de abajo.

Ilustración 11 Preparación de la librería en el entorno web de BaseSpace

Para posteriormente seleccionar la opción 2 (Libraries)

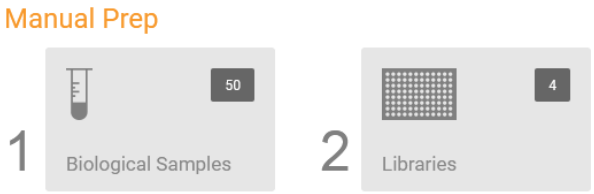


Ilustración 12 Selección de la librería en el entorno web de BaseSpace

En la página que se muestra haremos “click” en el botón de Import, eligiendo el fichero que hemos obtenido de iSkyLIMS.

Con estas acciones ya tenemos cargada los índices de las muestras de la carrera, que seleccionaremos para continuar con el proceso en BaseSpace

5.4.4 Inicio de la secuenciación

El paso siguiente es definir los POOLS de la carrera.

En la página que nos muestra BaseSpace iremos colocando las muestras en el POOL

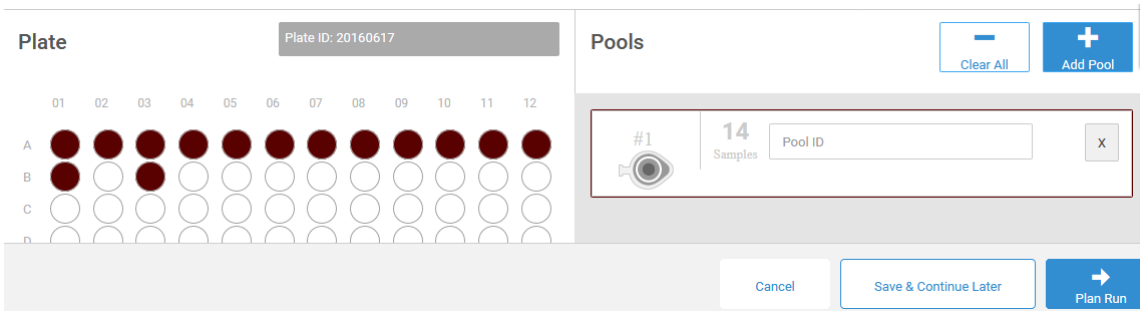


Ilustración 13 Definición de los “Pools” en el entorno web de BaseSpace

Una vez elegidas las muestras haremos click en el botón de “Plan Run”. Mostrando de nuevo otra página donde le indicaremos los parámetros como el tipo de instrumento, si vamos a utilizar “Single Read” o “Paired End”, el número de ciclos usados, etc.

Una vez completado este formulario daremos al botón de “Sequence para iniciar la secuenciación de la carrera”

Con esta última acción el secuenciador de illumina empezará su proceso de multiplicación del ADN.

5.4.5 Ficheros de salida

Dos tipos de ficheros van a generarse a través de iSkyLIMS. Un fichero de texto en formato csv que se creará como resultado de analizar el SampleSheet que el Wetlab Manager ha usado para definir la carrera en iSkyLIMS.

El formato de este fichero está formado por unas secciones de Header y Data.

[Header]

Compuesto por los siguientes campos:

FileVersion	**1**
LibraryPrepKit	**Nextera XT**
ContainerType	**PLATE**
ContainerID	**CONTAINER_ID**
Notes	**automatic generated file from iSkyLIMS**

Tabla 4 Campos incluidos en la sección de Header en el fichero a importar a illumina

[Data]

La sección de Data va a estar compuesto por una cabecera con los siguientes campos:

SampleID	Name	Species	Project	NucleicAcid	Well	Index1Name	Index1Sequence	Index2Name	Index2Sequence

Ilustración 14 Campos de la sección Data en el fichero a importa a illumina

A la que irán seguidas las filas de las muestras en el formato que precisa Base Space (https://basespace.illumina.com/) para importar el fichero en su sistema.

Este fichero estará disponible para su descarga desde 2 sitios diferentes. La primera vez cuando se ha generado el fichero y se muestra la página de resultados. Esta página no es accesible una vez que se ha salido de esta página, pero el fichero puede descargarse en cualquier momento buscando el nombre del proyecto, usando el icono de descarga.

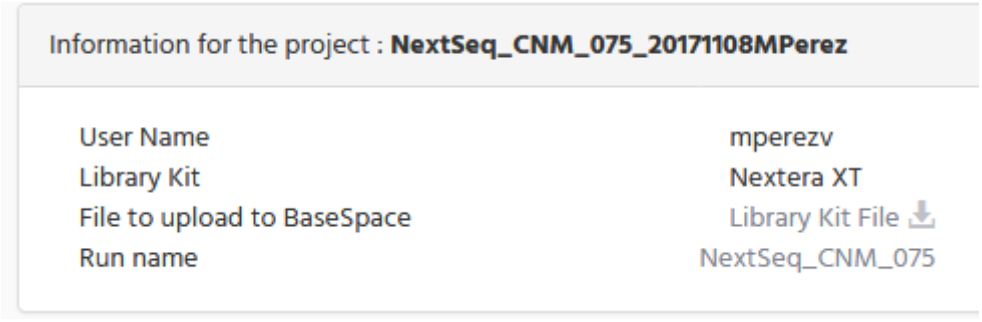


Ilustración 15 Pantalla de iSkyLIMS para la obtención del fichero que se importará e BaseSpace

5.5 Flujo en iSkyLIMS: Evolución de los estados de una carrera

Adicionalmente a los flujos de usuario y de secuenciación, que hemos comentado anteriormente, iSkyLIMS dispone de un flujo automático para completar el procesamiento de los ficheros de informes y estadísticas, que la conversión de BCL a fastq ha generado.

Para llevar a cabo esta tarea iSkyLIMS cuenta con la implementación de una máquina de estados, donde dependiendo del estado en que se encuentre se van a realizar un proceso determinado y exclusivo de su estado.

Durante este flujo la carrera irá pasando por los estados de SampleSent – Processing Running – Bcl2FastQ completed – Running Stats hasta llegar al estado Completed donde la información de las muestras y su calidad han sido procesadas y almacenadas en la base de datos de iSkyLIMS.

iSkyLIMS está orientado a gestionar los distintos estados de una carrera siguiendo la filosofía de máquina de estados, donde en función del estado donde se encuentre la carrera se realizarán determinadas acciones, no siendo posibles el resto de acciones que no son compatibles con su estado.

Como describimos anteriormente la creación de una carrera en iSkyLIMS empieza cuando el investigador ha completado el fichero SampleSheet.csv, que va a ser uno de los requisitos de entrada para poder iniciar el proceso de creación de la carrera.

Mientras el investigador está dentro del formulario de creación de carrera, esta se encuentra en un estado temporal llamado “PREPARED”.

Cuando el investigador ha rellenado satisfactoriamente las 2 partes del formulario del que componen la creación de la carrera en iSkyLIMS, se cambia al estado “RECORDED”, y guarda el fichero “sampleSheet.csv” en una carpeta, cuyo nombre es el índice de la carrera que ha sido asignado al almacenarlo en la base de datos.

A partir de este momento, se empiezan a generar 2 procesos automáticos que van a llevar la carrera al estado final de completado.

- El primer proceso automático que se ejecuta para una carrera es el “check_recorded_folder”
- El segundo proceso denominado “check_not_finish_run” se ejecutará una vez acabado el anterior

La figura siguiente muestra los estados por los que va a recorrer la carrera desde su creación hasta que está terminada.

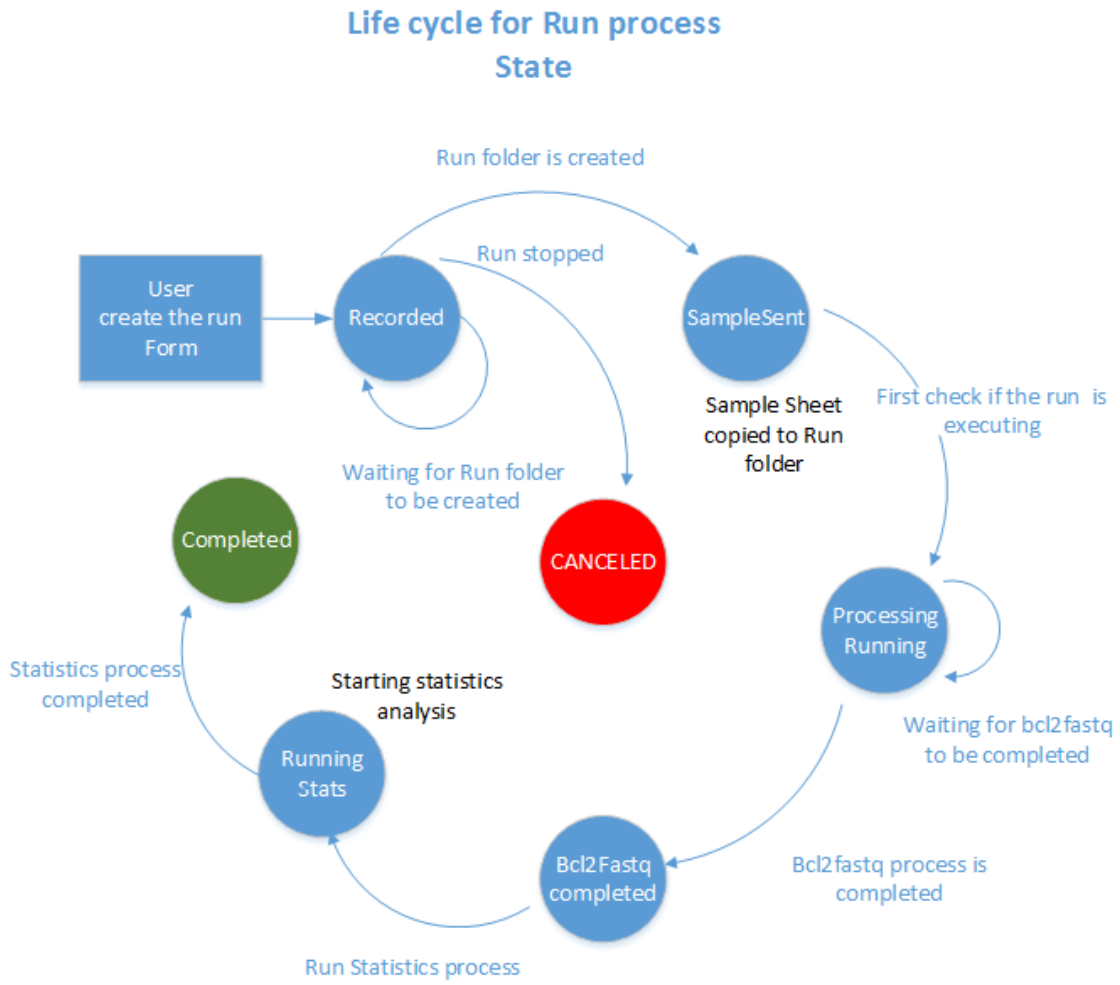


Ilustración 16 Diagrama de evolución de los estados por la que puede pasar una carrera

5.5.1 Evolución del estado “Recorded” a “SampleSent”

La carrera se encuentra en estado “Recorded” una vez que se ha completado satisfactoriamente el formulario de la carrera.

Para gestionar las carreras que están en “Recorded” se ha creado el proceso “check_recorded_folder” que se ejecuta a intervalos regulares definidos en el crontab del usuario de “django”.

Este proceso se va a encargar de poner el fichero “sampleSheet” en el servidor de almacenamiento de los datos de las carreras.

Para poder localizar la carpeta específica de la carrera, se va a abrir una conexión SAMBA entre el servidor donde corre iSkyLIMS y el servidor de almacenamiento y donde se van a ir recorriendo todas las carpetas para poder transferir el fichero “RunParameter.xml”. Una vez transferido al servidor de iSkyLIMS se analiza dicho fichero buscando la etiqueta de xml “ExperimentName”. Si el valor de este campo es igual al que se utilizó a la hora de rellenar el formulario.

Antes de copiar el fichero “SmapleSheet.csv” se comprueba que el estado de la secuenciación se ha completado. Para ello se tiene que leer la etiqueta “CompletionStatus” del fichero “RunCompletionStatus.xml”. Si el valor es “CompletedAsPlanned” significa que la secuenciación se ha terminado y los datos se han transferido al servidor de almacenamiento.

En este momento se copiará el fichero “SampleSheet.csv” en la carpeta del servidor, se borrará la carpeta temporal que se había creado para almacenar el fichero “SampleSheet.csv” y se cambiará el estado de la carrera a “SampleSent”.

En caso contrario de que no fuese igual el valor de la etiqueta de “ExperimentName” con el del formulario, se continuaría navegando por los directorios hasta encontrar la coincidencia.

Puede pasar que los ficheros de la carrera todavía no estuvieran copiados en el servidor y por tanto no va a ser posible la coincidencia. En este caso la carrera se mantendría en este estado esperando a que pase el tiempo definido en el crontab para la ejecución de nuevo del proceso “check_recorded_folder”.

Para optimizar este proceso se realizan los siguientes chequeos:

- Antes de abrir la conexión Samba al servidor de almacenamiento, se comprueba si existe alguna carrera que requiera ser procesada. Esto se hace comprobando si existen directorios donde están almacenados los “SampleSheet.csv”. En caso de que existan se pondría en marcha el proceso comentado anteriormente y si no el proceso termina sin necesidad de abrir ninguna conexión
- Para reducir el tiempo de recorrer todos los directorios para realizar la transferencia de los “RunParameter.xml” se ha creado un fichero “processed_run_file” en iSkyLIMS, donde cada vez que se encuentra una coincidencia (valor del ExperimentName con el nombre definido en el formulario), se actualiza este fichero añadiéndole el nombre del directorio. De esta forma, cuando se están navegando por los directorios para encontrar el Valor del ExperimetName que coincida con del formulario, se descartan la búsqueda en aquellos directorios que están definidos en el fichero, porque son de carreras que ya están procesadas.

5.5.2 Evolución del estado “Recorded” a “Canceled”

Una vez que se ha creado la carrera y mientras se está ejecutando el proceso en el secuenciador, el Investigador puede parar la ejecución. Si esto sucede el secuenciador lo indicará escribiendo en la etiqueta CompletionStatus el valor de “StoppedByUser” en el fichero RunCompletionStatus.xml.

Como comentábamos anteriormente una vez que se ha encontrado el directorio de la carrera, se va a comprobar si la ejecución del secuenciador se ha completado correctamente. Si no es así se borrarán los ficheros copiados en la carpeta temporal de iSkyLIMS, se eliminará el fichero de SampleSheet.csv, se añadirá en el fichero “processed_run_file” el directorio de esta carrera, se cambiará el estado “CANCELED” de la carrera y de los proyectos asociados a ella.

5.5.3 Evolución del estado “SampleSent” a “Processing Running”

Cuando se ha transferido el fichero “SampleSheet.csv” un script que está corriendo en el servidor está comprobando que todos los ficheros que ha genera el secuenciador de illumina y que el “SampleSheet.csv” se encuentra disponible. Con estas 2 condiciones el script empieza a ejecutar el comando “bcl2fastq” para convertir los ficheros bcl que ha generado el secuenciador al formato de Fastq, que será necesario para su posterior análisis.

Un segundo crontab se ha definido en el usuario de django para poder manejar el estado de las carreras desde que se ha enviado el SampleSheet.csv a la carpeta correspondiente y hasta que la carrera llega al estado “Complete”. La razón de tener 2 crontab separados es para poder definir tiempos diferentes, con el objetivo de que el proceso de copiar el SampleSheet.csv al servidor para que se puede realizar la conversión de BCL a fastq lo antes posible, con lo que el valor recomendado para este primer crontab se lanza cada media hora. Sin embargo, el segundo crontab puede ejecutarse con una periodicidad mayor porque el tiempo de conversión de BCL a fastq puede durar varias horas. Se recomienda poner la periodicidad de este crontab a 2 horas.

Cuando este segundo proceso “check_not_finish_run” se empieza a ejecutar va a interrogar a la base de datos de iSkyLIMS para coger el listado de las carreras que se encuentran en estado “Sample Sent”.

Con este listado el proceso irá comprobando si la conversión de BCL a fastq se ha completado. En la gran mayoría de los casos la conversión no ha terminado por lo que se va a actualizar el estado de la carrera a “Processing Running”, indicando en este estado de que la conversión todavía no ha finalizado.

5.5.4 Evolución del estado “Processing Running” a ”Bcl2FastQ completed”

El proceso “check_not_finish_run” al ejecutarse va a comprobar las carreras que se encuentran en estado “Preocessing Running”.

Para identificar que se ha realizado completamente la conversión se va a comprobar si se ha creado la carpeta Reports. Esta carpeta contiene parámetros de la calidad de las muestras y será la que usaremos posteriormente.

Una vez que se ha comprobado que la conversión ha finalizado, la carrera cambia al estado “Bcl2Fastq completed”

5.5.5 Evolución del estado “Bcl2FastQ completed”” a “Running Stats”

Una vez que la carrera se encuentra en el estado “Bcl2Fastq completed” se va a cambiar al estado “Running Stats”. Este estado es temporal (aproximadamente 2 minutos) y está creado con la intención de que no sea considerado a la hora de obtener estadísticas ni

información de la carrera debido a que durante el tiempo se están actualizando la información en la base de datos y dando lugar a información incompleta.

5.5.6 Evolución del estado “Running Stats”” a “Completed”

Cuando la carrera se encuentra en el estado “Running Stats” es cuando se van a analizar todos los ficheros de que se han generado en la conversión de BCL a fastq.

Durante este proceso, lo que se va a realizar primero es la copia de los ficheros, que se van a necesitar para la obtención de métricas de calidad.

Para evitar que se interrumpa la comunicación, entre iSkyLIMS y el servidor de almacenamiento, lo que primero se va a realizar es la copia de estos ficheros a una carpeta temporal de iSkyLIMS, donde una vez que se han transferido todos los ficheros se procederá a la extracción de la información.

Si durante esta transferencia de ficheros se pierde la conexión con el servidor de almacenamiento, se borrarán todas las copias de los ficheros que se hubieran transferido y se vuelve a cambiar el estado de la carrera a “Bcl2Fastq completed”, para que cuando nuevamente se vuelva a ejecutar el crontab entre esta carrera en la condición de ser procesada.

Una vez que todos los ficheros necesarios han sido transferidos se van analizar primero ConversionStats.xml y DemultiplexingStats.xml, donde vamos a obtener información de la calidad de las muestras. Esta información se va a almacenar en las siguientes tablas de la base de datos:

- **RawStatisticsXml**. Donde se van a almacenar los datos crudos que han sido parseados de los dos ficheros xml.
- **NextSeqStatsFlSummary**. Almacenará la información de los clusters, tanto en raw como filtrado, el número de bases y número de bases que ha obtenido para cada proyecto de la carrera
- **NextSeqStatsLaneSummary.** Se guardará información de los cluster, barcodes, número de bases, porcentaje del número de bases con una calidad mayor de 30 y la calidad media de las bases, para todas las muestras existentes en el proyecto.
- **RawTopUnknowBarcodes.** Va a contener la secuencia de las muestras que no han podido ser asignadas a ninguna de las muestras, el número de veces que se repiten, el número que le corresponde en el ranking (de 1 a 10) y la “lane” donde se ha encontrado dentro de la carrera.

Los siguientes ficheros que se van a analizar son los ficheros de datos (en formato binario) que están situados en la carpeta Interop. Para obtener información de estos ficheros utilizamos el paquete de Python de Interop. Este paquete nos va a permitir poder extraer información utilizando las funciones de este paquete.

Para guardar la información procesada se van a utilizar 3 tablas de la base de datos:

- **NextSeqStatsBinRunSummary**. Almacenando el número de muestras, el alineamiento, la tase de errores, número de muestras con calidad mayor de 30. Esta misma información se genera a 3 niveles diferentes; Total, Lane y Non Index

- **NextSeqStatsBinRunRead**. Va a almacenar el número de lecturas, el porcentaje de las muestras con una calidad mayor de 30, alineamiento, la tasa de errores, el número de muestras por Lane, para todas las Lane y los 4 Reads.
- **NextSeqGraphicsStats**. Va a contener la dirección donde se encuentran las seis gráficas que se obtienen por cada carrera:
    - ClusterCount
    - FlowCell
    - Intensity by cycle
    - Heat map
    - Histogram
    - SampleQc

Una vez que toda esta información esta guardada en el base de datos se cambia el estado del proyecto y de la carrera a “Completed”

5.6 Estructura de la base de datos

Para la creación de las tablas se ha utilizado el interfaz que tiene django cuando se definen los modelos. Django antepone el prefijo del nombre de la aplicación al nombre de la tabla, en nuestro caso la aplicación se llama “wetlab” por lo que todas las tablas wetlab_nombre del modelo.

Se han definido las siguientes tablas para almacenar toda la información que necesita iSkyLIMS:

- wetlab_runprocess;
- wetlab_nextseqstatsflsummary;
- wetlab_rawstatisticsxml;
- wetlab_rawstatisticsxml;
- wetlab_nextseqstatsflsummary;
- wetlab_rawtopunknowbarcodes;
- wetlab_nextseqstatsbinrunsummary;
- wetlab_nextseqstatsbinrunread;
- wetlab_nextseqgraphicsstats;
- wetlab_samplesinproject;
- wetlab_projects;
- wetlab_runningparameters;

La siguiente figura muestra la estructura de la base de datos de iSkyLIMS.

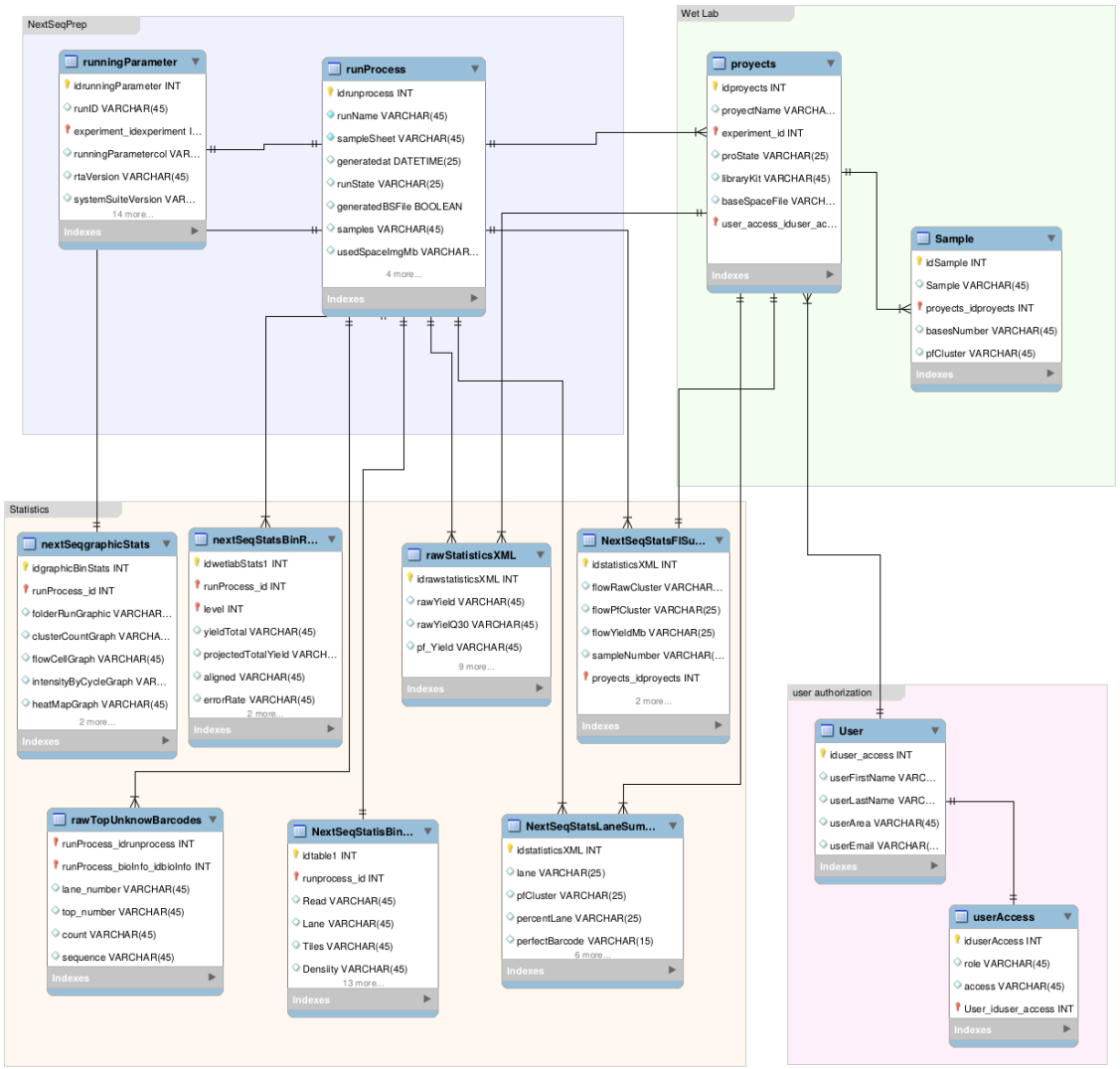


Ilustración 17 Diagrama de las tablas usadas en la base de datos de iSkyLIMS

La descripción de las tablas, así como la descripción de los campos que al componen están descritos en el capítulo de Anexo.

5.7 Logs

Para la creación de logs se utiliza el paquete de Python “logging”, que nos va a permitir además la creación de logs rotatorios para evitar que se llene el sistema con ficheros.

Esto nos da la seguridad de que nuestros ficheros nunca van a causar problemas de ocupación excesiva en el sistema, ya que como máximo la capacidad de los logs será la del tamaño que definamos que tiene cada fichero, multiplicado por el número de rotaciones. Lo que se ha definido para iSkyLIMS es un tamaño de fichero de 40Kb con una rotación de 5 ficheros.

Se van a utilizar dos logs, uno por cada proceso que se ejecutan desde el crontab

- check_recorded_folder.log
- checking_uncompleted_run.log

Cuando se ejecuta el crontab de cualquiera de estos 2 procesos se va a enviar la salida del terminal a otro log “crontab.log” donde se mostrará un texto indicando el inicio y final de cada llamada a los procesos, además de mostrar los errores que se produjeran al general el código.

5.8 2.5 Copias de seguridad

Las copias de seguridad que se van a realizar dentro de iSkyLIMS son a 3 niveles:

- **Backup de la base de datos.**

  Se ha creado un script para generar un backup diario de la base de datos de iSkyLIMS, transfiriendo el fichero comprimido a un servidor externo. En este servidor externo se ha creado otro proceso que mantiene el control del número de backups, borrando aquellos que tienen una antigüedad mayor de 1 mes.

- **Backup de los directorios de iSkyLIMS.**

  Otra información importante que tiene que ser preservada son los ficheros de entrada y salida que se han ido generando con el uso de iSkyLIMS. Un script que se ejecuta periódicamente y a la misma hora que el backup de la base de datos es generado. Para la realización de esta copia de seguridad vamos a hacer uso del programa rsync para tener siempre una copia actualizada de los ficheros.

- **Backup del entorno virtual.**

  El control y la ejecución de este backup será responsabilidad del departamento de sistemas, ya que son ellos los gestores de la máquina virtual donde corre iSkyLIMS. Este backup se realizará semanalmente.

5.9 Notificaciones a usuarios

ISkyLIMS está configurado para notificar al Wetlab Manager de que el procesamiento de los datos, obtenido de la conversión BCL a FastQ, han finalizado y están disponibles en iSkyLIMS. Para facilitar al Wetlab Manager la visualización de los resultados, se incluye dentro del cuerpo del correo, la URL de la carrera.

Adicionalmente se envía una copia al correo del administrador de iSkyLIMS, con efectos de seguimiento y control de la carrera.

5.10 Datos de la calidad de muestras y carreras

Como comentamos anteriormente, el programa que realiza la conversión de BCL a fastq, genera una serie de ficheros con datos de calidad de las muestras.

Los datos se van a mostrar en tres diferentes niveles:

- Nivel de muestra
- Nivel de proyecto
- Nivel de carrera

5.10.1 Datos mostrados a nivel de muestra

A nivel de estructura de la muestra se va mostrar:

- El nombre de la muestra
- El proyecto al que pertenece
- El nombre de la carrera.

Con el objetivo de poder identificar en que entorno se ha secuenciado la muestra.

En lo referente a los datos de la muestra se indican los siguientes parámetros:

- Barcode
- PF Cluster
- Porcentaje de la muestra en relación con la totalidad del proyecto
- Número de bases de la muestra
- Número de bases con una calidad mayor o igual a 30
- Calidad media de la muestra

Sample Quality Information

Sample	Barcode	PF Cluster	% of Project	Yield (Mbases)	>= Q30 bases	Mean Quality Score
AA-8361	GGACTCCT+TCTACTCT	2,740,389	1.66	906	83.029	32.313

Ilustración 18 Información de la calidad de la muestra

Se muestra además una gráfica para poder comparar las bases de la muestra contra todas las muestras realizadas en el proyecto.

Percentage of the Sample inside the project

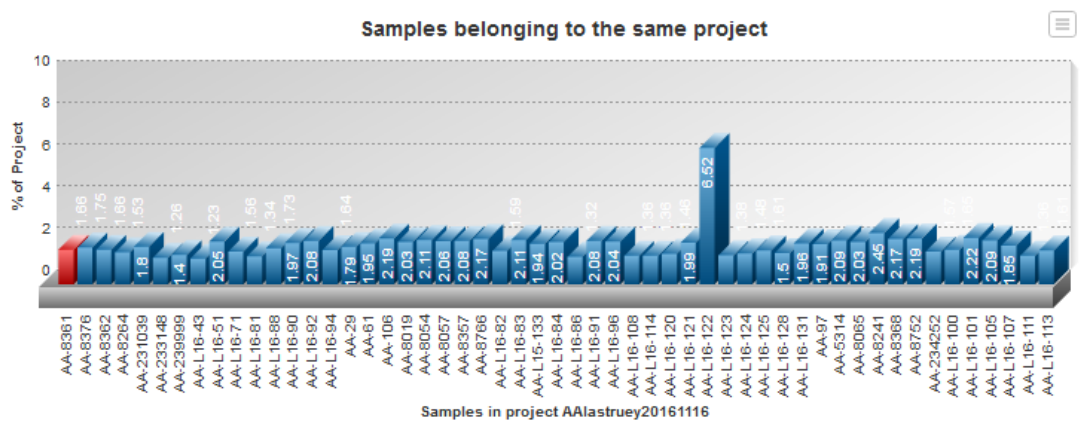


Ilustración 19 Comparativa de la muestra selecciona con el resto pertenecientes al mismo proyecto

5.10.2 Datos mostrados a nivel de proyecto

La información del proyecto está relacionada con el conjunto de muestras que pertenecen a dicho proyecto. Es decir, la información que mostrábamos a nivel de muestra, vamos a repetirla para cada una de las muestras.

Samples used for this project.

Sample	Barcode	PF Clusters	Percent of Project	Yield (Mbases)	% >= Q30 bases	Mean Quality Score
PAE1521	TAAGGCGA+ATAGAGAG	4,248,624	1.53	1,547	66.939	29.141
PAE1572	CGTACTAG+ATAGAGAG	4,254,543	1.54	1,551	67.587	29.277
PAE1450	AGGCAGAA+ATAGAGAG	4,212,699	1.52	1,553	67.382	29.231

Ilustración 20 Listado de las muestras que se han utilizado en el proyecto

Pero además se va a mostrar un resumen con el número de bases, número de muestras y el clúster.

Flowcell Summary for this project.

Cluster (Raw)	Cluster (PF)	Yield (MBases)	Number of Samples
173,946,158	164,699,827	52,529	53

Ilustración 21 Resumen general de la información del proyecto

Y otro resumen por cada “Lane”, mostrando además información de la calidad.

Lane Summary for this project.

Lane	PF Clusters	% of the lane	% Perfect barcode	% One mismatch barcode	Yield (Mbases)	% >= Q30 bases	Mean Quality Score
1	42,059,532	39.555	95.113	NaN	13,356	85.612	32.867
2	41,960,218	39.149	95.096	NaN	13,326	85.371	32.811
3	41,362,819	39.544	94.837	NaN	13,170	83.841	32.501
4	39,317,258	38.536	93.642	NaN	12,678	82.777	32.272

Ilustración 22 Resumen de la información de proyecto especificado por cada “Lane”

5.10.3 Datos mostrados a nivel de carrera

A nivel de carrera vamos a mostrar principalmente la información general, donde se van a incluir, qué proyectos forman la carrera, los parámetros de configuración de la misma y el fichero SampleSheet usado en la carrera, para poder descargarlo.

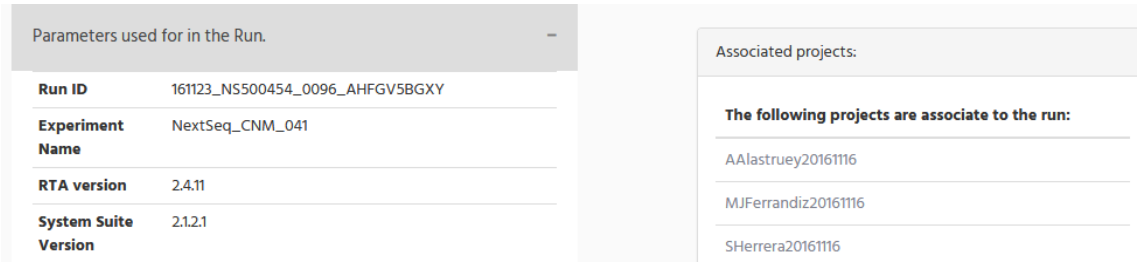


Ilustración 23 Parámetros utilizados en la ejecución de la carrera

El resumen de todos los proyectos indicando el número de bases, la calidad media de las muestras y las bases con una calidad mayor o igual a 30.

Las métricas de la carrera se van a indicar parámetros como el alineamiento, la tasa de errores e intensidad del ciclo 1.

Run Metrics | Lane Metrics | Charts

Run metrics

Level	Yield	Projected Yield	Aligned (%)	Error Rate (%)	Intensity Cycle 1	Quality >=30 (%)
Read 1	58.292	58.292	0.719	4.839	1498	67.243
Read 2	2.711	2.711	0.000	nan	1190	82.969
Read 3	2.714	2.714	0.000	nan	1147	80.593
Read 4	58.290	58.290	0.674	4.773	1781	60.960
Non Index	116.582	116.582	0.697	4.806	1640	64.102
Totals	**122.008**	**122.008**	**0.697**	**4.806**	**1404**	**64.888**

Ilustración 24 Información de la métricas de calidad de la carrera

En las métricas por Lane se van a indicar parámetros como la densidad, el número de lecturas, el alineamiento, la tasa de errores e intensidad del ciclo 1

Run Metrics | Lane Metrics | Charts

Lane Metrics

Read 1

Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	%>= Q30	Tield (G)	Cycles Err Rate	Aligned (%)	Error Rate (%
1	432	235 ± 14	67.885 ± 5.148	0.191 ± 0.1 / 0.147 ± 0.066	152.554	103.162	70.395	15.451	150	0.769 ± 0.082792	3.589 ± 0.922

Ilustración 25 Información por cada "Lane" de las métricas de calidad de las carreras

Durante el proceso de multiplexación de la muestra, algunos de los índices asociados a cada muestra pueden sufrir alguna degeneración, con lo que no va a ser posible identificar a qué muestra corresponden, formando parte de un grupo llamado "Unknow Barcodes" y que vamos a coger las 10 secuencias que más se repiten por cada Lane.

Se van a mostrar la cantidad de veces que la secuencia es encontrada en cada Lane y una gráfica para poder comparar las veces que aparecen las muestras en la carrera.

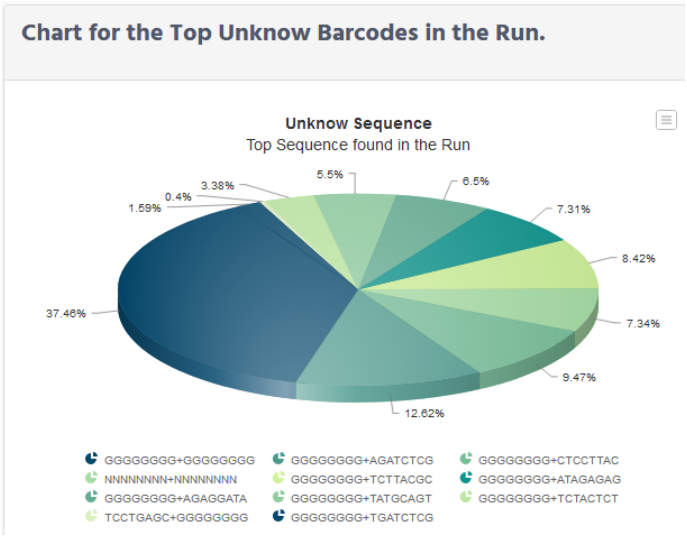


Ilustración 26 Comparativa de los "Unknown Barcodes" encontrados en la carrera

5.11 Estadísticas

Las estadísticas se han dividido en 4 niveles.

- Estadísticas por Investigador
- Estadísticas de las carreras realizadas durante un periodo de tiempo
- Estadísticas por tipo de experimento
- Estadísticas por kit de librería

5.11.1 Estadísticas por Investigador

Las estadísticas de investigador están relacionadas con los proyectos que ha realizado, durante un periodo que es seleccionado mediante el formulario.

Como resultado de la búsqueda se mostrarán los nombres de proyectos mostrándose gráficas de calidad (Q>30 y calidad media) para poder comparar todos estos proyectos entre sí. Estas gráficas están divididas por Lane.

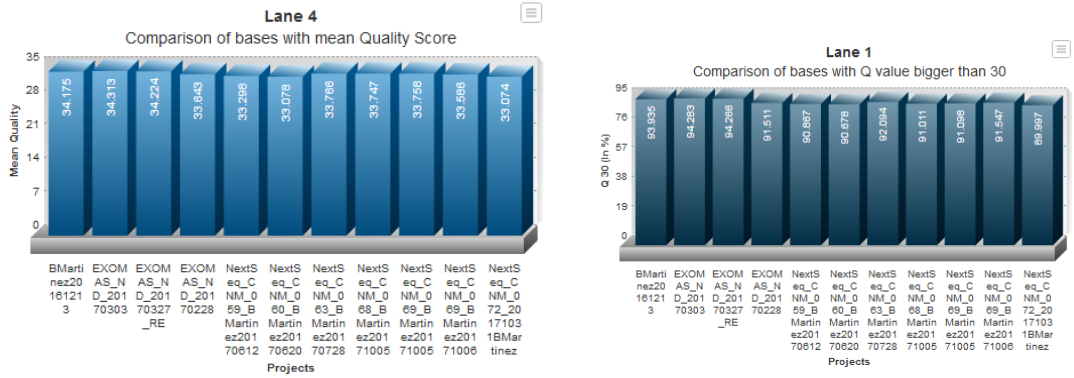


Ilustración 27 Comparativa de las estadísticas del investigador en todos sus proyectos

Como parte de las estadísticas por investigador se van a mostrar también unas gráficas de calidad (Q> 30 y calidad media) comparándolas con el resto de proyectos.

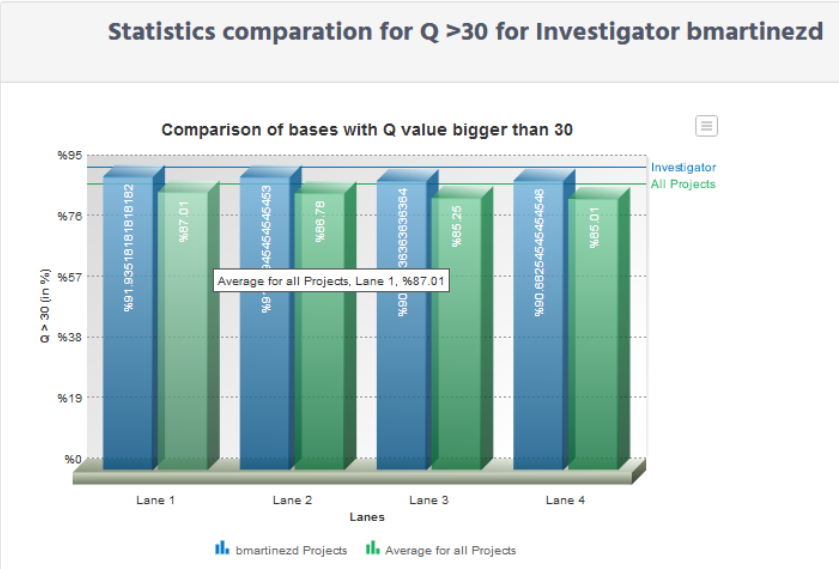


Ilustración 28 Comparativa del investigador con el resto de proyectos

5.11.2 Estadísticas de las carreras realizadas durante un periodo de tiempo

En muchos casos nos va a interesar ver las carreras que se han realizado durante un periodo en concreto, por ejemplo: el último mes, el último trimestre, o durante unas semanas en particular. Para poder tener esta flexibilidad se pide al usuario que introduzca la fecha de inicio y la de final.

Para este periodo de tiempo se van a mostrar el nombre de las carreras que se han realizado.

Además de los datos de calidad de estas carreras, se van a mostrar las secuencias de los índices “Undetermined” y el número de veces que se repiten por cada Lane.

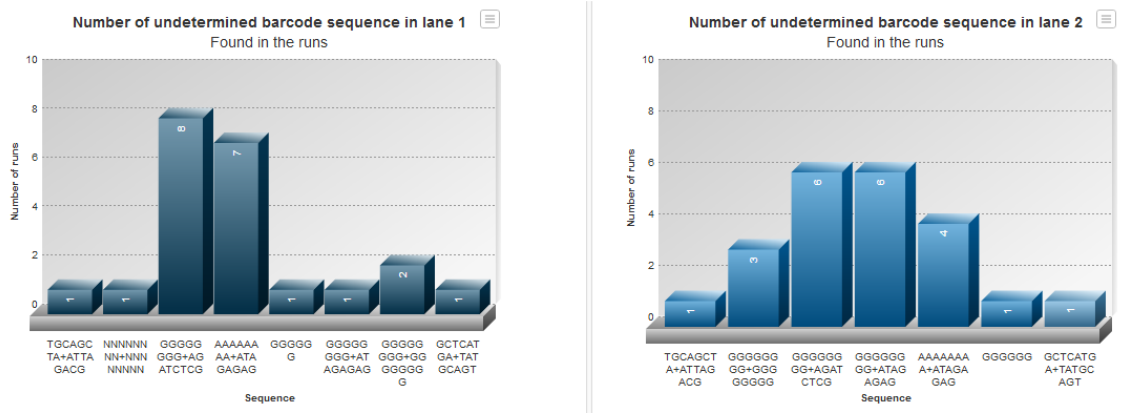


Ilustración 29 Comparativa de las carreras realizadas durante un periodo de tiempo

5.11.3 Estadísticas por kit de librería

Cuando realizamos la búsqueda del kit de librería podemos tener 2 posibles escenarios:

- Que sólo se encuentre un tipo de kit de librería
- Que más de un tipo de librería es encontrado

En el caso de que sólo se encuentre un tipo de librería, se van a mostrar todos los proyectos que se han encontrado para el periodo de tiempo seleccionado.

Se mostrarán gráficas del número de bases, la calidad media y la calidad de Q> 30 de cada proyecto y separada por Lane.

Para poder comparar este kit de Librería con el resto se van a mostrar gráficas de porcentaje de Bases de Q>30, Calidad media y el número de bases.

En el caso de que se encuentre más de un tipo de librería se van a mostrar los datos de Número de bases, calidad media y porcentaje de Q>30 de estas bases para poder compararlas.

Se podrá también comparar estos kits de Librería con el resto mostrándose gráficas de porcentaje de Bases de Q>30, Calidad media y el número de bases.

5.12 Informes

La generación de informes se va a realizar en 3 tramos de tiempo:

- **Informes Anuales.**

  Recoge la información de las carreras que se han creado desde enero a diciembre.

- **Informes trimestrales**

  Recoge la información de las carreras que se han creado en uno de los cuatrimestres del año, es decir:

  - Primer Cuatrimestre. Desde enero a marzo
  - Segundo Cuatrimestre. Desde abril a junio
  - Tercer Cuatrimestre. Desde julio a septiembre
  - Cuarto Cuatrimestre. Desde octubre a diciembre

- **Informes mensuales**

  Se recoge la información del día 1 hasta el último día de ese mes.

Para todos estos informes se mostrará en primer lugar el nombre de las carreras que se han realizado en este periodo e indicando las carreras que se han completado y las que están pendientes de completarse.

Una carrera que no está terminada, puede deberse a varias razones:

- El proceso de secuenciación no ha finalizado todavía

- Ha habido algún problema en el momento de transferirse los ficheros desde el secuenciador hasta el servidor de almacenamiento
- El proceso ha sido abortado por el wetlab manager

Se van a mostrar estadísticas del número de proyectos que han realizado los investigadores. Se han dividido en 3 grupos:

- Investigadores que han realizado menos de 5 Proyectos en el año
- Investigadores que han realizado entre 5 y 10 proyectos
- Investigadores que han realizado más de 10 proyectos

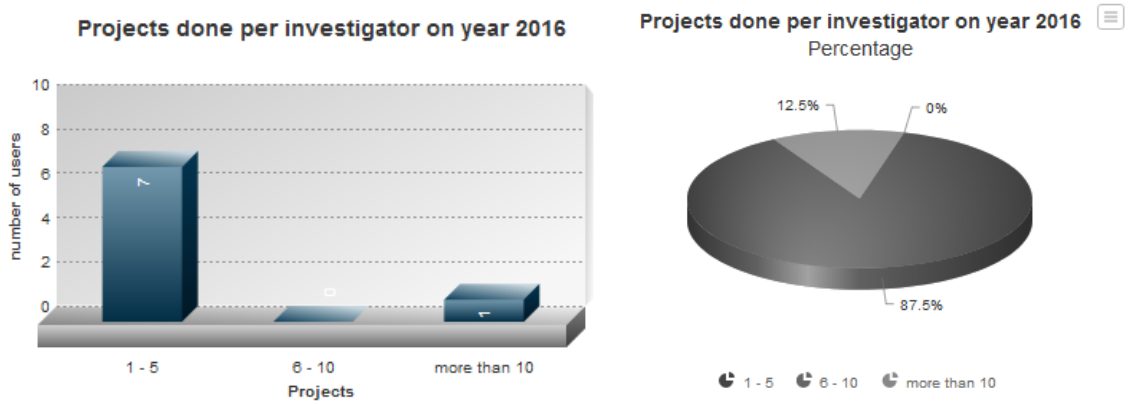


Ilustración 30 Informe mostrando los proyectos realizados por cada investigador por grupo

Se mostrará las comparativas de Alineamiento, Tasa de errores, Calidad Q>30 de las carreras realizadas durante el periodo que se ha seleccionado a la hora de realizar el informe

Ilustración 31 Comparativa en el informe de la calidad Q> 30

5.13 Sistema empleado para la implementación

Para la implementación del LIMS se han empleado 2 sistemas:

Pre-Producción: Que es donde se ha ido haciendo el desarrollo y se han realizado pruebas para verificar su funcionamiento.

Producción. Es el entorno donde se va a hacer público a los usuarios.

El sistema de pre-producción está montado sobre una máquina virtual desplegada sobre VMCenter6.0 sobre un servidor (HP Proliant D385 G7, AMD Opteron de 16 núcleos (x2), 128GB RAM, 8TB tipo SAS) ubicado en el Centro de Procesamiento de Datos (CPD) del ISCIII.

El sistema de producción estará montado sobre una máquina virtual desplegada sobre VMCenter6.0 sobre un servidor (HP Proliant D385 G7, AMD Opteron de 16 núcleos (x2), 128GB RAM, 8TB tipo SAS) igualmente ubicado en el Centro de Procesamiento de Datos (CPD) del ISCIII.

En ambos sistemas (pre-producción y producción) correrá sobre un sistema operativo CentOS 6.9 donde está configurado un servidor Apache con el módulo WGSI para suministrar las páginas web.

A la hora de elegir el framework sobre el cuál correría la aplicación web se estuvieron evaluando en 2 diferentes entornos:

- Basado en PHP
- Basado en Python

La principal ventaja de elegir un entorno basado en PHP es la madurez del lenguaje y la gran variedad de diferentes Frameworks que están disponibles, de los que podríamos seleccionar aquellos que tuvieran unas características más acordes a nuestros requerimientos, alta disponibilidad, comunidad de uso y soporte, modularidad, etc. Entre todos ellos seleccionamos 3: “Symfony”, Laverel, o Yii

Si bien en PHP podríamos disponer de un gran número de opciones, en Python este número era menor, y de entre ellas nos quedamos con Django, Flask, y Web2py.

Llegados a este punto todo apuntaba a utilizar un entorno basado en PHP, pero lo que nos hizo decantarnos por el uso de Python es que estábamos desarrollando una aplicación bioinformática y que esta tendría que hacer uso de paquetes específicos de biología.

Al elegir definitivamente Python como entorno, se optó por Django, por la gran comunidad de usuarios que tiene, cada vez es más popular, donde la lista de sitios (Pinterest, Instagram, Onion, etc) donde se está usando Django no para de crecer, dando con ello una fiabilidad en su uso.

6 Discusión

Actualmente, en muchos laboratorios, los procedimientos para el seguimiento y el almacenamiento de la información de la muestra se basan en hojas de cálculo (Excel) o en pequeñas bases de datos personales (Access), sin una verdadera gestión o estandarización de la información.

Esta clase de gestión individualizada de datos tiende a generar ineficiencias y redundancias, que dan lugar a un aumento de los potencialmente errores (típicamente errores tipográficos) que son difíciles de rastrear y/o resolver.

Disponer de un LIMS que maneje y centralice las actividades del laboratorio es una necesidad básica que todo laboratorio dedicado a la secuenciación masiva debería de implementar si quiere tener un control y seguimiento preciso de las muestras que procesa.

Existen bastantes sistemas LIMS disponibles, bien sean de pago o bien “open source”, sin embargo, cualquiera de estas soluciones va a requerir un desarrollo adicional para poder adaptarlas a las necesidades específicas de cada laboratorio.

En nuestro caso particular, partimos de que a la hora de realizar este proyecto no existe ningún LIMS que sea open source que maneje muestras provenientes del sistema de NextSeq de illumina, ni tampoco que se adaptase a los requerimientos que se debían cubrir. Se analizó la posibilidad de modificar alguna de las soluciones existentes, pero se vio que el tiempo de analizar en detalle la solución, añadir el nuevo código, así como eliminar aquellas partes que no fueran de utilidad para nuestras necesidades era incluso mayor que el de partir de una base completamente limpia. Por lo que se optó finalmente por crear una nueva solución orientada a dar solución a las necesidades del laboratorio del iSCIII.

iSkyLIMS está basado en un entorno WEB para facilitar la interacción con el usuario, permitiendo aislar la complejidad del procesamiento de los datos, con el objetivo de que pueda ser utilizado por los investigadores independientemente de sus conocimientos informáticos.

La ventaja de utilizar el framework de Django es que por una parte nos facilita la integración de los demás programas realizados en Python para el procesamiento de los ficheros y por otra parte nos evita tener que re-escribir un código para mostrar las páginas Web, el manejo de sesiones y el registro de los usuarios que se han “loggeado”.

Cualquier LIMS que definamos ha de confiar en una base de datos para llevar el control de todos los pasos y almacenar de forma fiable la información. En nuestro caso hemos utilizado MySQL como base de datos debido a su escalabilidad y flexibilidad, así como la de tener una amplia comunidad de usuarios para su soporte, asegurándonos que la base de datos será actualizada ante el descubrimiento de fallos en la seguridad o bugs.

En muchos proyectos se enfatiza tanto en definir una buena solución, que nos satisfaga tanto del punto de vista visual como el de resultado que se suele obviar los aspectos de seguridad de la aplicación. Nos estamos refiriendo al uso de logs, y copias de seguridad.

La creación de LOGS es necesaria para poder verificar que está pasando, para poder determinar de esta manera si es un fallo del programa o se debe a otra situación, para realizar con ello las medidas correctivas en el código. Tenemos que tener presente que una gran parte del código de iSkyLIMS se realiza de forma transparente al usuario y por tanto debe mantenerse constancia si la ejecución de los procesos han sido debidos a la ejecución programada a través del “crontab” o se debe a otra causa.

Las copias de seguridad es algo imprescindible y que en ocasiones nos solemos acordar de ellas cuando ya es demasiado tarde y hemos perdido gran parte o toda la información/código. Para evitar cualquier eventualidad que se pudiera producir iSkyLIMS está configurado para realizar copias de seguridad a 3 niveles:

- A nivel de la base de datos.
- A nivel de código y de ficheros
- A nivel de máquina virtual

Por ello si existe un problema con la base de datos podremos recuperar la información del día anterior. Si por error se han borrado ficheros podemos recuperarlo de la última copia de seguridad. Si existiera algún problema con el entorno de VMCenter, como la corrupción de algún disco que estuviera utilizando iSkyLIMS, también podremos solventar esta eventualidad recuperando de la copia de seguridad.

Una de las principales ventajas de iSkyLIMS es la de poder mostrar estadísticas a distintos niveles, (muestras, proyectos y carreras) adaptándose de esta forma a los distintos roles definidos en iSkyLIMS.

La comparación de proyectos por investigador nos da información en 2 áreas.

Por un lado, nos muestra la calidad de las muestras de los proyectos en los que ha participado el investigador permitiendo al investigador comprobar si todos sus proyectos mantienen una calidad constante y la de descubrir a simple vista si algún proyecto de los realizados tiene una calidad inferior

Y por otro lado va a permitir determinar si hay diferencias de calidad entre los diferentes procedimientos y métodos que realizara un investigador con respecto al resto de investigadores.

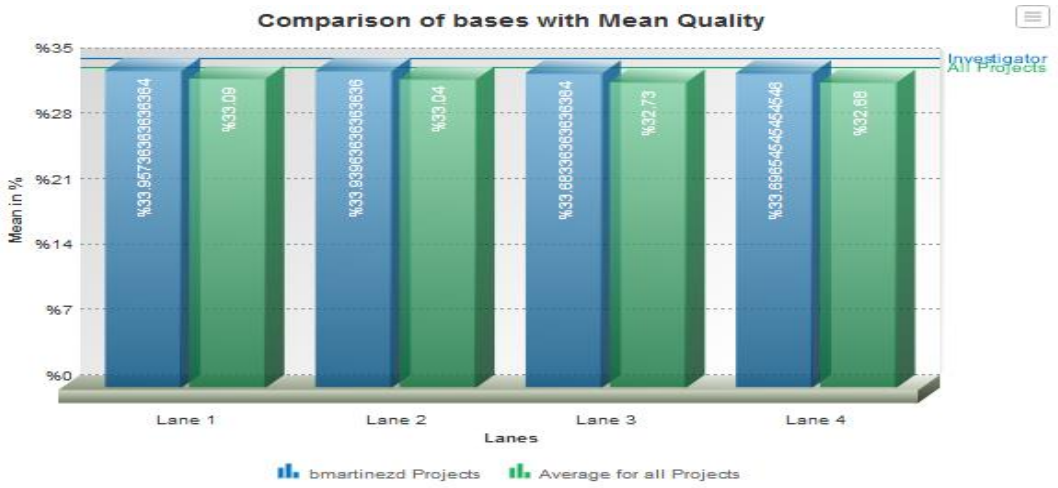


Ilustración 32 Comparativa de la calidad media del investigador con el resto de investigadores

Mostrando la comparación entre los kits de librerías nos va a permitir identificar aquellos kits que nos proporcionan mejor calidad en las muestras, y también lo opuesto, identificar aquellos kits que dan unas peores lecturas comparadas con el resto de kits.

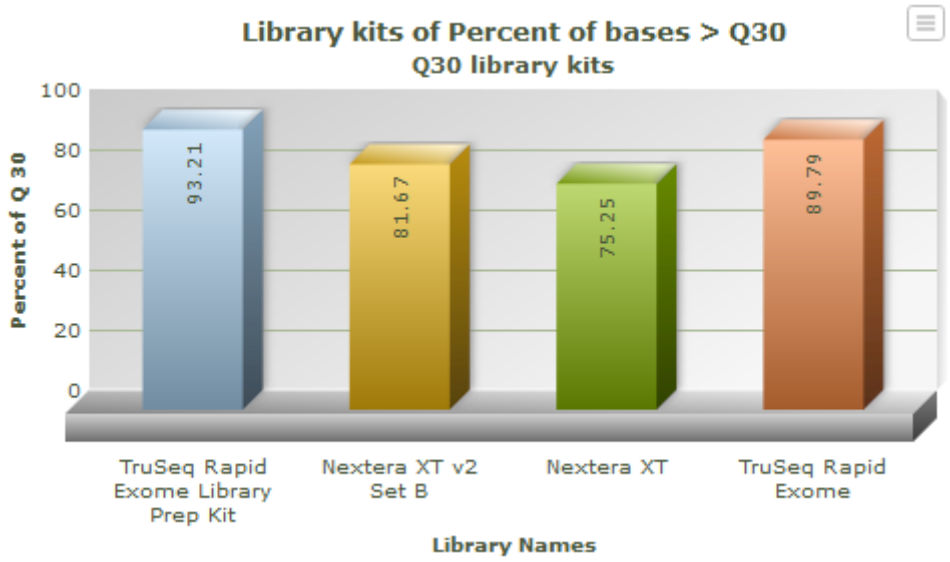


Ilustración 33 Comparativa de la calidad Q> 30 de los kits de librería

Otra medida importante es la determinación de la cantidad de "Unknow barcodes" que se han producido en la carrera. Con ello podremos determinar que secuencias son las que mayormente se degeneran en el proceso de la multipexación con el objetivo de intentar reducir en la medida de lo posible su utilización.

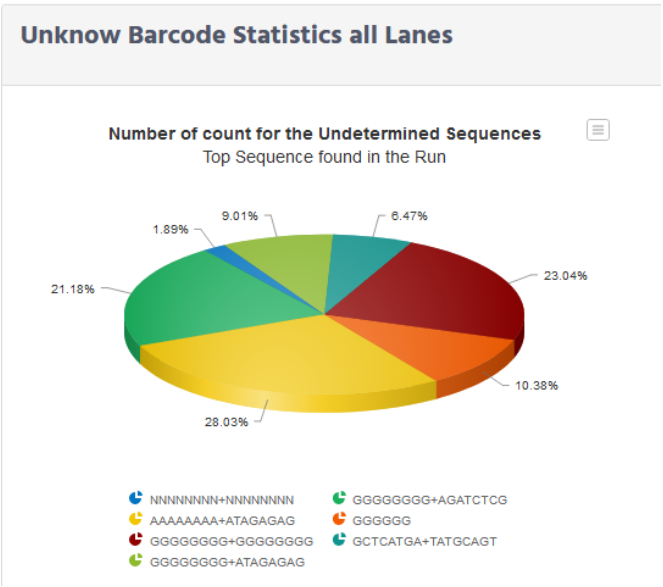


Ilustración 34 Comparativa de los "Unknow Barcodes" de las carreras

Todas las estadísticas que proporciona iSkyLIMS van a servir para poder verificar la calidad de las muestras, proporcionando datos para realizar un estudio pormenorizado que conlleve a descubrir fallos en la calidad de las muestras.

Un LIMS no estaría completo si no estuviera preparada para obtener informes que permitan optimizar los procedimientos de laboratorio.

La creación de los informes está orientada hacia los Wetlab Manager, para que puedan generan informes con carácter; mensual, trimestral o anual, de las carreras y proyectos realizados.

7 Conclusiones

El uso de un iSkyLIMS permite

1.- Eludir el control artesanal de hojas de cálculo o la administración local de archivos, para centrarse en la automatización de los procedimientos de laboratorio para crear un estándar en el seguimiento de muestras, y la automatización derivada de NGS.

2.- Monitorear el flujo de los datos y estandarizar los directorios de entrada y salida y los nombres de los archivos, incluso cuando se usan múltiples protocolos de análisis en los mismos datos.

3.- Asegurar la trazabilidad completa del análisis realizado, así como el almacenamiento de los datos.

4.- Permitir a los usuarios no experimentados ejecutar análisis a través de una interfaz gráfica de usuario (GUI) que actúa como front-end para las canalizaciones.

5.- La monitorización de procesos facilitando al laboratorio aumentar la eficiencia y reducir los errores manuales, presentando bases concretas para mantener la escalabilidad de los proyectos.

8 Bibliografía

1. Web Development with Django Cookbook. Ed. Packt Publishing (16 de octubre de 2014)
2. Venco et al.: SMITH: a LIMS for handling next- generation sequencing workflows. BMC Bioinformatics 2014 15(Suppl 14): S3.SMITH: a LIMS for handling next-generation sequencing workflows
3. Illumina BaseSpace: [https://basespace.illumina.com/home/index].
4. Genologics Clarity LIMS. [http://www.genologics.com/claritylims].
5. Nix DA, Di Sera TL, Dalley BK, Milash BA, Cundick RM, Quinn KS, Courdy SJ:
6. Scholtalbers J, Rossler J, Sorn P, de Graaf J, Boisguerin V, Castle J, Sahin U: Galaxy LIMS for next-generation sequencing. Bioinformatics 2013, 29:1233-1234.
7. Clarity LIMS. [http://www.genologics.com/claritylims]
8. Van Rossum T, Tripp B, Daley D: SLIMS–a user-friendly sample operations and inventory management system for genotyping labs. Bioinformatics 2010, 26(14):1808–1810.
9. Grimes et al: MendeLIMS: a web-based laboratory information management system for clinical genome sequencing. BMC Bioinformatics 2014, 15:290
10. Django project: [https://www.djangoproject.com/]
11. Mysql : [https://www.mysql.com/]

9 Anexos

9.1 Descripción de la estructura de la base de datos

Para una mejor comprensión de las tablas, se describen estas tablas dentro de la fase donde son utilizadas.

9.1.1 Tablas de la definición de la carrera

Para guardar la información de la carrera iSkyLIMS va a utilizar las siguientes tablas:

- wetlab_runprocess
- wetlab_projects

En la tabla **wetlab_runprocess** se va a almacenar la información relacionada con la carrera y va a contener estos campos:

Nombre	Descripción	Tipo de campo
runName	Se va a guardar en este campo el nombre del experimento que es introducido en el formulario de la definición de la carrera	CharField (45)
sampleSheet	Va a contener el nombre del fichero que el wetlab manager ha usado en la definición de la carrera.	FileField
runState	Indicará el estado en el que está la carrera	CharField (25)
samples	Contiene el número de muestras que contiene la carrera	CharField (45)
requestCenter	Va a almacenar el centro responsable de la carrera	CharField (45)
generatdat	Guardará la fecha y la hora de cuando se ha grabado la carrera en iSkyLIMS	DateTimeField

Tabla Suplementaria 1 Descripción de tabla wetlabrunprocess

La tabla **wetlab_projects** guardará información de los proyectos que componen la carrera y constará de los siguientes campos:

Nombre	Descripción	Tipo de campo
runprocess_id	Guarda el id de la relación con la carrera a la que pertenece el proyecto	ForeignKey
projectName	Almacena el nombre del proyecto	CharField (45)
procState	Estado por los que pasa el proyecto hasta que se finaliza la obtención de resultados	CharField (25)
libraryKit	El nombre de la librería usada en el proyecto	CharField (45)
baseSpaceFile	Dirección donde se guarda el fichero para ser importado al BaseSpace de illumina	CharField (255)
generatdat	Guardará la fecha y la hora de cuando se ha grabado el proyecto en iSkyLIMS	DateTimeField

Tabla Suplementaria 2 Descripción de tabla wetlab_projects

9.1.2 Tablas para almacenar los datos de configuración de la carrera

Durante la extracción de los datos de configuración de la carrera se van a usar estas tablas:

- wetlab_runprocess
- wetlab_projects
- wetlab_runningParameter

Las tablas de wetlab_runprocess y wetlab_projects actualizarán el campo del estaco modificándolo al estado “SampleSheet sent”.

La tabla **wetlab_runningParameter** va a almacenar los datos de como se ha configurado la carrera conteniendo los siguientes campos:

Nombre	Descripción	Tipo de campo
runName_id	Clave para enlazar con la tabla RunProcess	Primary Key

RunID	Nombre de la carpeta que asigna BaseSpace al generar los ficheros de la carrera.	CharField (255)
RunName	Nombre de la carrera que fue definido en el formulario a la hora de subir el fichero de Sample Sheet	CharField (255)
RTAVersion	Versión de la RTA	CharField (255)
SystemSuiteVersion	Version de Software de la Suite de illumina	CharField (255)
LibraryID	Identificador de la librería usada en la carrera	CharField (255)
Chemistry	Tipo de secuenciador utilizado en la carrera	CharField (255)
RunStartDate	Fecha de cuando se ha ejecutado la carrera	CharField (255)
AnalysisWorkflowType	Tipo de análisis	CharField (255)
RunManagementType		CharField (255)
PlannedRead1Cycles	Ciclo de lecturas para el Read 1	CharField (255)
PlannedRead2Cycles	Ciclo de lecturas para el Read 2	CharField (255)
PlannedIndex1ReadCycles	Ciclo de lecturas para el índice 1	CharField (255)
PlannedIndex2ReadCycles	Ciclo de lecturas para el índice 2	CharField (255)
ApplicationVersion	Versión del aplicativo	CharField (255)
NumTilesPerSwath		CharField (255)
ImageChannel	Canales de imagen utilizados	CharField (255)
Flowcell	Identificador del Flowcell utilizado	CharField (255)
ImageDimensions	Dimensiones de la imagen	CharField (255)
FlowcellLayout		CharField (255)

Tabla Suplementaria 3 Descripción de tabla wetlab_runningParameters

9.1.3 Tablas para coger datos en la obtención de las estadísticas

La mayoría de las tablas de iSkyLIMS van a guardar datos de la calidad de la carrera para ser utilizada El nivel de estadísticas va a contener las tablas donde se van a guardar los datos de la carrera y que servirán de base para crear las gráficas y las tablas de estadísticas. Para esta fase van a verse afectadas las siguientes tablas

- wetlab_RawStatisticsXml
- wetlab_RawTopUnknowBarcodes

- wetlab_NextSeqStatsFlSummary
- wetlab_NextSeqStatsLaneSummary
- wetlab_NextSeqStatsBinRunSummary
- wetlab_NextSeqStatsBinRunRead
- wetlab_NextSeqGraphicsStats.

Las tablas de wetlab_runprocess y wetlab_projects actualizarán el campo del estado modificándolo al estado “completed” una vez acabada la incorporación de datos en las tablas anteriormente mencionadas.

La tabla **wetlab_rawstatisticsxml** va a almacenar los datos crudos que se han obtenido a la hora de pasear los ficheros ConversionStats.xml y DemultiplexingStats.xml. La información incluida es a nivel de muestras que componen la carrera.

Esta tabla va a tener relaciones 1:n con la tabla wetlab_runProcess y con wetlab_projects. Va a contener los siguientes campos:

Nombre	Descripción	Tipo de campo
runprocess_id	Relación 1:n con la tabla RunProcess	ForeignKey
project_id	Relación 1:n con la tabla Projects	ForeignKey, null=True
defaultAll	Va a identificar si los datos son de todos los proyectos o el “default” que contendrá las muestras que no ha podido relacionar con ningún index.	CharField (40) null=True
rawYield	Número de bases de la muestra	CharField (255)
rawYieldQ30	Número de bases con una calidad mayor de 30	CharField (255)
rawQuality	Calidad de la muestra	CharField (255)
PF_Yield	Número de bases de la muestra aplicando el filtro	CharField (255)
PF_YieldQ30	Número de bases con una calidad mayor de 30 aplicando el filtro	CharField (255)
PF_QualityScore	Calidad de la muestra aplicando el filtro	CharField (255)
generated_at	Fecha y hora de cuando se ha almacenado el record en la base de datos	DateTimeField

Tabla Suplementaria 4 Descripción de tabla wetlab_rawstatisticsxml

La tabla **wetlab_RawTopUnknowBarcodes** va a almacenar las secuencias que no han podido ser asociadas a ninguna muestra porque el “barcode” no coincide con el que se han definido y por tanto no puede ser asignado a ninguna de ellas.

Esta información se ha obtenido a la hora de pasear el fichero **DemultiplexingStats.xml**. Esta tabla va a tener una relación 1:n con wetlab_runProcess. Va a contener los siguientes campos:

Nombre	Descripción	Tipo de campo
runprocess_id	Relación 1:n con la tabla RunProcess	Primary Key
lane_number	Número de Lane	, null=True
top_number	Ranking de posición en función del número de secuencias encontradas	CharField (40)
count	Número de secuencias encontradas para una misma secuencia	CharField (40)
sequence	Secuencia	CharField (40)

generated_at	Fecha y hora de cuando se ha almacenado el record en la base de datos	DateTimeField

Tabla Suplementaria 5 Descripción de tabla wetlab_RawTopUnknowBarcodes

La tabla **wetlab_NextSeqStatsFlSummary** va a almacenar los datos procesados de la tabla wetlab_RawStatisticsXml creando, en la base de datos, una fila por cada proyecto existente en la carrera.

Cada carrera contendrá información de la suma de todos los proyectos y otra que contendrá los índices que no se han podido asignar. Esta tabla va a tener una relación 1:n con wetlab_runProcess. Va a contener los siguientes campos:

Nombre	Descripción	Tipo de campo
runprocess_id	Relación 1:n con la tabla RunProcess	ForeignKey
lane_number	Relación 1:n con la tabla Projects	ForeignKey, null=True
defaultAll	Va a identificar si los datos son de todos los proyectos o el "default" que contendrá las muestras que no ha podido relacionar con ningún index.	CharField (40) null=True
flowRawCluster		CharField (40)
flowPfCluster		CharField (40)
flowYieldMb		CharField (40)
sampleNumber	Número de muestras utilizadas en el projecto	CharField (40)
generated_at	Fecha y hora de cuando se ha almacenado el record en la base de datos	DateTimeField

Tabla Suplementaria 6 Descripción de tabla wetlab_NextSeqStatsFlSummary

La tabla **wetlab_NextSeqStatsLaneSummary** va a almacenar los datos procesados de la tabla wetlab_RawStatisticsXml creando, en la base de datos, una fila por cada lane y por cada proyecto existente en la carrera. Cada carrera contendrá información por lane de la suma de todos los proyectos y la información por "lane" de los índices que no se han podido asignar. Esta tabla va a tener una relación 1:n con RunProcess. Va a contener los siguientes campos:

Nombre	Descripción	Tipo de campo
runprocess_id	Relación 1:n con la tabla RunProcess	ForeignKey
lane_number	Relación 1:n con la tabla Projects	ForeignKey, null=True
defaultAll	Va a identificar si los datos son de todos los proyectos o el "default" que contendrá las muestras que no ha podido relacionar con ningún index.	CharField (40) null=True
lane		CharField (40)
pfCluster		CharField (64)
percentLane		CharField (64)
perfectBarcode		CharField (64)
oneMismatch		CharField (64)
yieldMb		CharField (64)
biggerQ30		CharField (64)
meanQuality		CharField (64)

generated_at	Fecha y hora de cuando se ha almacenado el record en la base de datos	DateTimeField

Tabla Suplementaria 7 Descripción de tabla wetlab_NextSeqStatsLaneSummary

La tabla **wetlab_NextSeqStatsBinRunSummary** va a almacenar los datos procesados de los ficheros binarios localizados en la carpeta de "interop" creando, en la base de datos, una fila por cada "read", otra línea para los índices que no se han podido asignar y la última el total de todas las filas anteriores.

Esta tabla va a tener una relación 1:n con wetlab_runProcess. Va a contener los siguientes campos:

Nombre	Descripción	Tipo de campo
runprocess_id	Relación 1:n con la tabla RunProcess	ForeignKey
level	Va a identificar el número de Read (1, 2, 3, 4) o si es la información de los índices sin asignar o la información total de la suma de los Reads	CharField (10)
yieldTotal	Número de bases secuenciadas	CharField (10)
projectedTotalYield	Número de bases esperadas ser secuenciadas	CharField (10)
aligned	El porcentaje de la muestra que alinea con el genoma PhiX	CharField (10)
errorRate	El cálculo de la tasa de error al alinear con el genoma PhiX	CharField (10)
intensityCycle	El promedio de la intensidad del canal A medido en el primer ciclo promediado sobre los grupos filtrados	CharField (10)
biggerQ30	La estadística de intensidad correspondiente en el ciclo 20 como un porcentaje de ese valor en el primer ciclo. 100% x (Intensidad en el ciclo 20) / (Intensidad en el ciclo 1).	CharField (10)

Tabla Suplementaria 8 Descripción de tabla wetlab_NextSeqStatsBinRunSummary

La tabla **wetlab_NextSeqStatsBinRunRead** va a almacenar los datos procesados de los ficheros binarios localizados en la carpeta de "interop" creando, en la base de datos, una fila por cada Read y por cada lane.

Esta tabla va a tener una relación 1:n con wetlab_RunProcess. Va a contener los siguientes campos:

Nombre	Descripción	Tipo de campo
runprocess_id	Relación 1:n con la tabla RunProcess	ForeignKey
read	Va a identificar el número de Read (1, 2, 3, 4)	CharField (10)
level	Va a identificar el número de Lane (1, 2, 3, 4)	CharField (10)
tiles	Número de Tiles por Lane	CharField (40)
density	La densidad de clusters (en miles por mm2) detectada por análisis de imagen, +/- una desviación estándar.	CharField (40)
cluster_PF	Porcentaje de clusters que han pasado el ciclo con +/- una desviación estándar	CharField (40)

phas_prephas	El valor utilizado por RTA para el porcentaje de moléculas en un grupo para el cual la secuenciación se queda atrás (phasing) o salta adelante (prephasing) el ciclo actual dentro de una lectura.	CharField (40)
reads	Número de clusters (en Millones)	CharField (40)
reads_PF	Número de clusters que han pasado el filtro (en Millones)	CharField (40)
q30	La estadística de intensidad correspondiente en el ciclo 20 como un porcentaje de ese valor en el primer ciclo. 100% x (Intensidad en el ciclo 20) / (Intensidad en el ciclo 1).	CharField (40)
yields	Número de bases que han pasado el filtro	CharField (40)
cyclesErrRated	Número de ciclos con tasa de error al alinear con el genoma PhiX	CharField (40)
aligned	El porcentaje de la muestra que alinea con el genoma PhiX	CharField (40)
errorRate	El cálculo de la tasa de error al alinear con el genoma PhiX	CharField (40)
errorRate35	Tasa de error para ciclos del 1 al 35	CharField (40)
errorRate50	Tasa de error para ciclos del 1 al 50	CharField (40)
errorRate75	Tasa de error para ciclos del 1 al 75	CharField (40)
errorRate100	Tasa de error para ciclos del 1 al 100	CharField (40)
intensityCycle	El promedio de la intensidad del canal A medido en el primer ciclo promediado sobre los grupos filtrados	CharField (40)

Tabla Suplementaria 9 Descripción de tabla wetlab_NextSeqStatsBinRunRead

La tabla **wetlab_NextSeqGraphicsStats** va a almacenar la dirección donde están guardados las gráficas generadas al ejecutar los comandos de obtención de gráficas, usando los ficheros binarios localizados en la carpeta de “interop”.

Esta tabla va a tener una relación 1:1 con RunProcess. Va a contener los siguientes campos:

Nombre	Descripción	Tipo de campo
runprocess_id	Relación 1:1 con la tabla RunProcess	ForeignKey
folderRunGraphic	Directorio donde se van a almacenar los ficheros de las gráficas	CharField (255)
cluserCountGraph	Nombre del fichero para la gráfica de cluserCountGraph	CharField (255)
flowCellGraph	Nombre del fichero para la gráfica de flowCellGraph	CharField (255)
intensityByCycleGraph	Nombre del fichero para la gráfica de intensityByCycleGraph	CharField (255)
heatMapGraph	Nombre del fichero para la gráfica de heatMapGraph	CharField (255)
histogramGraph	Nombre del fichero para la gráfica de histogramGraph	CharField (255)
sampleQcGraph	Nombre del fichero para la gráfica de sampleQcGraph	CharField (255)
generated_at	Fecha y hora de cuando se ha almacenado el record en la base de datos	DateTimeField

9.2 Referencias a las Ilustraciones

9.3 Referencias a las Tablas

9.4 Referencias a las Tablas Suplementarias